Principal Component Estimation of a Large Covariance Matrix with High-Frequency Data*

Yacine Aït-Sahalia†
Department of Economics
Princeton University and NBER

Dacheng Xiu‡
Booth School of Business
University of Chicago

This Version: September 21, 2015

Abstract

Under a large dimensional approximate factor model for asset returns, we use high frequency data to infer their covariance structure. We adapt principal component analysis (PCA) to this high frequency setting and provide an asymptotic theory that covers joint in-fill time series and diverging cross-sectional dimension asymptotics, under a variety of sparsity assumptions on the idiosyncratic covariance matrix. Empirically, we investigate the factor structure of a large portfolio of stock returns, focusing in particular on the consistency of the latent factor structure with their counterparts based on well-known observable factors in the literature.

Keywords: high-dimensional data, high-frequency latent factor model, PCA, Global Industrial Classification Standard (GICS), low rank plus sparse, S&P 500 index constituents

1 Introduction

The celebrated arbitrage pricing theory by Ross (1976) suggests that assets earn risk premia because they are exposed to underlying risk factors, and that the co-movement between assets are driven by their exposure to these risk factors. However, the factor structure of asset returns is either neglected or abused by some portfolio managers in practice. For example, the sample covariance matrix is often used as the key input to portfolio optimization, which ignores the factor structure behind asset returns, leading to economically infeasible portfolios in particular when the universe of assets under management is enormous. At the other extreme, some users build factor models with tens

*We are benefited from extensive discussions with Jianqing Fan, Alex Furger, Chris Hansen, Jean Jacod, Yuan Liao, Markus Pelger, and Weichen Wang, as well as seminar and conference participants at CEMFI, the 8th Annual SoFiE Conference, 2015 IMS-China International Conference on Statistics and Probability, and the 11th World Congress of the Econometric Society. We thank Chaoxing Dai for excellent research assistance.

†Address: 26 Prospect Avenue, Princeton, NJ 08540, USA. E-mail address: yacine@princeton.edu.

‡Address: 5807 S Woodlawn Avenue, Chicago, IL 60637, USA. E-mail address: dacheng.xiu@chicagobooth.edu.

or hundreds of factors, which defeats the purpose of the factor investment, not to mention many of such factors are spurious.

The aforementioned problems are partially due to the lack of theoretical guidance on the number of factors or the choice of factors to use, when building large covariance matrices. Additionally, a small number of factors alone cannot fully explain the comovement of asset returns. In the universe of equities, individual firms' characteristics may be responsible to their comovement. One notable example is the dual of Pepsi and Coca-Cola, whose correlation cannot be fully explained by their exposure to systematic factors. Using more systematic factors is not an ideal solution to these problems.

Using the principal component analysis (PCA), we separate correlations of high-frequency returns driven by the common yet latent factors. We then find a striking block-diagonal pattern in the residual correlations, after sorting the stocks by their firms' global industrial classification standard (GICS) codes. This finding suggests that the covariance matrix can be approximated by a low-rank component due to their exposure to some common factors, plus a sparse component, which reflects their sector/industry specific exposure.

We thereby take this structure into account when estimating the covariance matrix. Our analysis is based on a general continuous-time semiparametric approximate factor model, allowing the well-documented time-variation in equity volatilities and correlations. We show that both the factor-driven and the residual components of the sample covariance matrix are identifiable, as the cross-sectional dimension increases. Our PCA based estimator is not only consistent, but also invertible and well-conditioned. Additionally, based on the eigenvalues of the sample covariance matrix, we provide a new estimator for the number of latent factors. To develop the theoretical properties of these estimators, and in particular to highlight the effect of an increasing dimensionality, we use the joint in-fill and diverging dimensionality asymptotics.

Empirically, we find that the factors uncovered from the PCA explain more variation of asset returns than that explained by observable portfolio factors such as the market portfolio, the Fama-French portfolios, as well as the industrial ETF portfolios. The difference is not very large when industrial factors are included, suggesting that the industrial portfolios perhaps span approximately the same linear space as the estimated latent factors. Also, the residual covariance matrix based on the PCA is more sparse than that based on observable factors, and both demonstrate a clear block-diagonal pattern. Finally, we take the new covariance estimator to an empirical test-drive, and find that both estimators perform significantly better against the sample covariance estimator in an out-of-sample portfolio allocation setting.

There has been a large literature on the factor model and its applications in finance and macro-economics. The arbitrage pricing theory by Ross (1976) and the ICAPM by Merton (1973) develop the fundamental economic theory behind the factor structure of asset returns. Chamberlain and Rothschild (1983) extend Ross' strict factor model to the approximate factor model, in which the residual covariances are not necessarily diagonal, hence allowing the comovement unaccountable from the systematic risk factors. Based on this model, Bai and Ng (2002) propose a statistical method-

ology to determine the number of factors, and Bai (2003) studies the statistical inference of factors and their loadings. All these papers use the principal component analysis (PCA), which has been adopted by Connor and Korajczyk (1988) to test the arbitrage pricing theory, and by Stock and Watson (2002) in a forecasting setting. All these papers assume latent factors. On the other hand, many efforts have been devoted to search for observable empirical proxies of the latent factors. Among many others, the Fama-French 3-factor model by Fama and French (1993) is perhaps most widely used. Their factors are explicitly constructed using portfolios formed by sorting firm characteristics. Chen, Roll, and Ross (1986) consider macroeconomic variables as factors, for example, inflation, output growth gap, interest rate, risk premia, and term premia.

The above factor models are static, as opposed to the dynamic factor models discussed in Forni, Hallin, Lippi, and Reichlin (2000), Forni and Lippi (2001), and Forni, Hallin, Lippi, and Reichlin (2004), in which the lagged values of the unobserved factors may also affect the observed dependent variables. Both static and dynamic factor models are cast in discrete time. In contrast, our paper discusses continuous-time factor models, where the observed variables are continuous Itô semimartingales. Our setting is particularly suitable for analyzing stock returns observed within a fixed window. Prior literature in this setting mainly discusses regression models with observable explanatory variables. For example, Mykland and Zhang (2006) introduce the ANOVA as well as the univariate regression. Todorov and Bollerslev (2010) add a jump component to their univariate regression setting. Aït-Sahalia, Kalnina, and Xiu (2014) extend their model to allow multivariate regressors and time-varying coefficients. An exception is by Aït-Sahalia and Xiu (2014), which introduces the nonparametric PCA for Itô semimartingales, sheding light on the latent factor structure of asset returns. This paper, however, imposes a semiparametric factor model and adopts the PCA to estimate the covariance and residual covariance matrices.

With respect to the literature on large covariance matrix estimation, Fan, Fan, and Lv (2008) propose an estimator based on observable factors using a strict factor model. Fan, Liao, and Mincheva (2011) study the approximate factor model, introducing high-dimensional thresholding techniques to the residual covariance matrix. Closely related to our estimator is the POET estimator proposed by Fan, Liao, and Mincheva (2013). They adopt the same low-rank plus sparsity structure of the approximate factor model. Related papers also include Fan and Wang (2014) and Fan, Liao, and Wang (2014). Alternative estimators include the shrinkage approach by Ledoit and Wolf (2004a), Ledoit and Wolf (2004b), and Ledoit and Wolf (2012); the thresholding approach by Bickel and Levina (2008a), Bickel and Levina (2008b), Cai and Liu (2011), etc. Zhou, Cai, and Ren (2014) provides a comprehensive summary of the literature.

Our paper is also related to the growing literature on the covariance matrix estimation with high-frequency data. The vast amount of data available intraday makes it rather attractive to estimate the comovement between assets nonparamtrically, as opposed to building complex parametric models with years of daily data. Earlier literature mainly focuses on attacking the microstructure noise and asynchronous observation issues endemic to multivariate high-frequency data, which result in a significant bias in the sample covariance matrix estimates by Barndorff-Nielsen and Shephard (2004).

Among others, Aït-Sahalia, Fan, and Xiu (2010), Christensen, Kinnebrock, and Podolskij (2010), Barndorff-Nielsen, Hansen, Lunde, and Shephard (2011), Zhang (2011), and Bibinger, Hautsch, Malec, and Reiß (2014) propose different noise-robust estimators to the synchronized data using either the previous tick scheme or the refresh time method. Shephard and Xiu (2012) propose a rate-efficient estimator that further achieves the desired positive-semidefinite property.

However, when the dimension of the asset universe increases to a few hundreds, the number of observations of the synchronized data drops dramatically, so that the curse of dimensionality becomes the dominant problem that plagues the covariance estimation. Fan, Furger, and Xiu (2015) establish the consistency of a noise-robust estimator, allowing the dual in-fill and diverging dimensionality asymptotics. Related work also includes Tao, Wang, Yao, and Zou (2011), Tao, Wang, and Zhou (2013), and Tao, Wang, and Chen (2013). All these papers share the assumption that the population covariance matrix itself is sparse. This assumption is refuted by Fan, Furger, and Xiu (2015), which then propose a regression approach that decomposes the covariance matrix into a low-rank component driven by some observable factors, and a sparse component that reflects the residual correlations. They use up to 12 factors, including the three Fama-French factors as well as 9 sector SDPR Exchange Traded Funds (EFTs). In contrast, our estimator uses the PCA approach to separate the low rank component and the sparse components apart, which accommodates latent factors. Independently, Pelger (2015a) derives the distributional theory for factors and their loadings in a high-frequency factor model with jumps, and proposes an alternative estimator for the number of factors. The distributional theory therein is entry-wise, similar to the results developed in Bai and Ng (2002) and Bai (2003), whereas our paper discusses the matrix-wise asymptotic properties for the covariance matrix and its inverse. See Pelger (2015b) for some related empirical work.

The structure of the rest of the paper is as follows. Section 2 sets up the model and provides the assumptions. Section 3 details the econometric analysis that provides the theoretical support for our procedure. Section 4 provides Monte Carlo simulation evidence. Section 5 includes an empirical study that demonstrates the performance of our estimator. Section 6 concludes. The appendix contains mathematical proofs.

2 Model Setup and Assumptions

Let $(\Omega, \mathcal{F}, \{\mathcal{F}_t\}, \mathbb{P})$ be a filtered probability space. Let $\mathcal{M}_{d\times r}$ be the Euclidian space of $d \times r$ matrices, and $\mathcal{M}^{++}_{d\times r}$, $\mathcal{M}^{+}_{d\times r}$ be the subsets of positive-definite and positive-semidefinite matrices, respectively. Throughout the paper, we use $\lambda_j(A)$, $\lambda_{\min}(A)$, and $\lambda_{\max}(A)$ to denote the jth, the minimum, and the maximum eigenvalues of a matrix A. In addition, we use $\|A\|_1$, $\|A\|$, and $\|A\|_{\mathrm{F}}$ to denote the $\mathbb{L}_1$ norm, the operator norm (or $\mathbb{L}_2$ norm), and the Frobenius norm of a matrix A, that is, $\max_{i,j}\sum_i |A_{ij}|$, $\sqrt{\lambda_{\max}(A^\intercal A)}$, and $\sqrt{\mathrm{Tr}(A^\intercal A)}$, respectively. When A is a vector, both $\|A\|$ and $\|A\|_{\mathrm{F}}$ are equal to its Euclidean norm. We also use $\|A\|_{\mathrm{MAX}} = \max_{i,j}|A_{ij}|$ to denote the $\mathbb{L}_\infty$ norm of A on the vector space. We use e_i to denote a d-dimensional column vector whose ith entry is 1 and 0 elsewhere. K is a generic constant that may change from line to line.

We observe a large intraday panel of asset prices, Y, at $0, \Delta_n, 2\Delta_n, \ldots, t$, where Δ_n is the sampling frequency. We assume Y follows a continuous-time factor model,

$$Y_t = \beta X_t + Z_t, \tag{1}$$

where Y_t is a d-dimensional vector process, X_t is a r-dimensional *unobservable factor* process, Z_t is the idiosyncratic component, and β is a constant factor loading matrix of size $d \times r$. The constant β assumption is not foreign to the literature. In fact, Reiß Todorov, and Tauchen (2015) find supportive evidence of this assumption using high-frequency data.

To complete the specification, we make additional assumptions on the dynamics of factors and the idiosyncratic components.

Assumption 1. *Suppose the vector of log asset prices Y follows a factor model given by (1), in which X and Z are continuous Itô semimartingales, that is,*

$$X_t = \int_0^t h_s\, ds + \int_0^t \eta_s dW_s, \quad Z_t = \int_0^t f_s ds + \int_0^t \gamma_s dB_s.$$

We denote the spot covariance of X_t as $e_t = \eta_t \eta_t^\intercal$, and that of Z_t as $g_t = \gamma_t \gamma_t^\intercal$. W_t and B_t are independent Brownian motions. In addition, h_t and f_t are progressively measurable, the process η_t, γ_t are càdlàg, and e_t, e_{t-}, g_t, and g_{t-} are positive-definite. Finally, for all $1 \leq i, j \leq r$, $1 \leq k, l \leq d$, there exist a constant K and a locally bounded process H_t, such that $|\beta_{kj}| \leq K$, and that $|h_{i,s}|$, $|\eta_{ij,s}|$, $|\gamma_{kl,s}|$, $|e_{ij,s}|$, $|f_{kl,s}|$, and $|g_{kl,s}|$ are all bounded by H_s for all ω and $0 \leq s \leq t$.

Apart from the fact that jumps are excluded, Assumption 1 is fairly general, allowing almost arbitrary forms of heteroscedasticity in both X and Z. While jumps are potentially important to explain asset return dynamics, we do not find them essential in large-scale portfolio allocation exercises. We thereby leave jumps aside for future work. The assumption on the uniform bounds of all processes is necessary to develop the large dimensional asymptotic results. This is a fairly standard assumption in the factor model literature, e.g., Bai (2003).

We also impose the usual exogeneity assumption. Different from those discrete-time regression or factor models, this assumption imposes path-wise restrictions, which is standard in a continuous-time factor model.

Assumption 2. *For any $1 \leq j \leq r$, $1 \leq k \leq d$, and $0 \leq s \leq t$, $[Z_{k,s}, X_{j,s}] = 0$, where $[\cdot, \cdot]$ denotes the quadratic covariation.*

Combing with Equation (1), Assumptions 1 and 2 determine a factor structure on the spot covariance matrix of Y, denoted as c_s:

$$c_s = \beta e_s \beta^\intercal + g_s. \quad 0 \leq s \leq t,$$

This leads to a key equality:

$$\Sigma = \beta \mathrm{E} \beta^\intercal + \Gamma, \tag{2}$$

where, without ambiguity we omit the dependence of Σ, E, and Γ on t,

$$\Sigma = \frac{1}{t}\int_0^t c_s ds, \quad \Gamma = \frac{1}{t}\int_0^t g_s ds, \quad \text{and} \quad \mathrm{E} = \frac{1}{t}\int_0^t e_s ds.$$

Throughout the paper, t is fixed to be 1 month. The number of factors r is unknown but finite, whereas d increases to ∞ as Δ goes to 0.

Finally, we impose some structure on the residual covariance matrix Γ.

Assumption 3. *Γ is a block diagonal matrix, and the set of its non-zero entries, denoted by S, is known prior to the estimation. Moreover, $\lambda_{\min}(\Gamma)$ is bounded away from 0 almost surely.*

The block-diagonal assumption on Γ is motivated from our empirical work as well as a closely-related study by Fan, Furger, and Xiu (2015). Fan, Furger, and Xiu (2015) find such a pattern of Γ in their regression setting, after sorting the stocks by the GICS code and stripping off the part explained by observable factors. Our empirical study sheds light on an even more clear block-diagonal pattern. The comparison between the two studies reinforce the low-rank plus sparsity structure behind asset returns.

To ensure the identification of Γ, we need to control the size of the largest block in Γ, which is given by

$$m_d = \max_{1\le i\le d} \sum_{1\le j\le d} 1_{\{\Gamma_{ij}\neq 0\}}.$$

This quantity coincides with the commonly used measure of the sparsity of a matrix. For example, Bickel and Levina (2008a) introduce this notion of sparsity to the covariance matrix, and establish the asymptotic theory of a thresholded sample covariance matrix estimator. The degree of sparsity determines the convergence rate of the estimator. We impose the sparsity assumption on the residual covariance matrix, because this low-rank plus sparsity structure matches the asset returns we have.

In a setting with low-frequency time series data, Fan, Liao, and Mincheva (2011) and Fan, Liao, and Mincheva (2013) use the sparsity assumption without assuming the block-diagonal pattern of Γ. We make this additional assumption to achieve the simplicity of the estimator. The next section provides a simple nonparametric covariance matrix estimator, with easy-to-interpret tuning parameters, the number of digits of the GICS code and the number of latent factors. We also provide a new estimator to determine the number of factors.

3 Econometric Analysis

3.1 Identification

There is fundamental indeterminacy in a latent factor model. For instance, we can rotate the factors and their loadings simultaneously without changing the covariance matrix Σ. In the literature, the canonical form of a factor model assumes that the covariance matrix E is an identity matrix and $\beta^\intercal\beta$ is diagonal. This is not appropriate for our model, since we allow the factor covariance matrix

E to be time-varying and non-deterministic, which is more general than the common factor models in discrete-time.

Our goal in this paper is to propose a new covariance matrix estimator, taking advantage of the assumed low-rank plus sparsity structure. We do not, however, try to identify the factors or their loadings, which can be pinned down by imposing sufficiently many identification restrictions, see, e.g., Bai and Ng (2013). Since we only need to separate $\beta \mathrm{E} \beta^\intercal$ and Γ from Σ, we can avoid some strict and unnecessary restrictions.

Chamberlain and Rothschild (1983) study the identification problem of a general approximate factor model. One of their key identification assumptions is that the eigenvalues of Γ are bounded, whereas the eigenvalues of $\beta \mathrm{E} \beta^\intercal$ diverge because the factors are assumed pervasive. It turns out that we can relax the boundedness assumption on the eigenvalues of Γ. In fact, the block-diagonal structure on Γ, combined with some sparsity condition imposed on m_d, implies that the largest eigenvalue of Γ diverges but at a slower rate compared to the eigenvalues of $\beta \mathrm{E} \beta^\intercal$.

These considerations motivate our identification assumption below, which is weaker than the usual canonical-form and pervasiveness assumptions in the literature.

Assumption 4. E *is a positive-definite covariance matrix, with distinct eigenvalues bounded away from 0. Moreover,* $\left\| d^{-1} \beta^\intercal \beta - \mathrm{I}_r \right\| = o(1)$*, as* $d \to \infty$*.*

This leads to our identification result.

Theorem 1. *Suppose Assumptions 3 and 4 hold. Also, assume that* $\|\mathrm{E}\|_{\mathrm{MAX}} \leq K$*,* $\|\Gamma\|_{\mathrm{MAX}} \leq K$ *almost surely, and that* $d^{-1/2} m_d = o(1)$*. Then* r*,* $\beta \mathrm{E} \beta^\intercal$*, and* Γ *can be identified as* $d \to \infty$*. That is,* $\bar{r} = r$*, if* d *is sufficiently large. Moreover, we have*

$$\left\| \sum_{j=1}^{\bar{r}} \lambda_j \xi_j \xi_j^\intercal - \beta \mathrm{E} \beta^\intercal \right\|_{\mathrm{MAX}} \leq K d^{-1/2} m_d, \quad \text{and} \quad \left\| \sum_{j=\bar{r}+1}^{d} \lambda_j \xi_j \xi_j^\intercal - \Gamma \right\|_{\mathrm{MAX}} \leq K d^{-1/2} m_d,$$

where $\{\lambda_j, 1 \leq j \leq d\}$ *and* $\{\xi_j, 1 \leq j \leq d\}$ *are the eigenvalues and their corresponding eigenvectors of* Σ*, and* $\bar{r} = \arg\min_{1 \leq j \leq d} (d^{-1} \lambda_j + j d^{-1/2} m_d) - 1$*.*

The key identification condition is $d^{-1/2} m_d = o(1)$, which creates a sufficiently wide gap between two groups of eigenvalues, so that we can identify the the number of factors as well as the two components of Σ. The identification is only possible when d is sufficiently large – the so called "the blessings of dimensionality." This is in contrast with the result for a classical strict factor model, where the identification is achieved by matching the number of equations with the number of unknown parameters.

3.2 Estimation Procedure

To fix ideas, let $\Delta_i^n X = X_{i\Delta_n} - X_{(i-1)\Delta_n}$, for $1 \leq i \leq n = [t/\Delta_n]$. Our estimator is built on the principal component analysis of the sample covariance matrix estimator. Denote

$$\widehat{\Sigma} = \frac{1}{t} \sum_{i=1}^{n} (\Delta_i^n X)(\Delta_i^n X)^\intercal.$$

Suppose that $\widehat{\lambda}_1 > \widehat{\lambda}_2 > \ldots > \widehat{\lambda}_d$ are the simple eigenvalues of $\widehat{\Sigma}$, and that $\widehat{\xi}_1, \widehat{\xi}_2, \ldots, \widehat{\xi}_d$ are the corresponding eigenvectors.[1] Our covariance matrix estimator $\widehat{\Sigma}^S$ is given by

$$\widehat{\Sigma}^S = \sum_{j=1}^{\widehat{r}} \widehat{\lambda}_j \widehat{\xi}_j \widehat{\xi}_j^\intercal + \widehat{\Gamma}^S, \tag{3}$$

where $\widehat{r}$ is an estimator of r discussed below,

$$\widehat{\Gamma} = \sum_{j=\widehat{r}+1}^{d} \widehat{\lambda}_j \widehat{\xi}_j \widehat{\xi}_j^\intercal, \quad \text{and} \quad \widehat{\Gamma}^S = (\widehat{\Gamma}_{ij} 1_{(i,j)\in S}), \tag{4}$$

The residual covariance matrix estimator $\widehat{\Gamma}^S$ is a by-product.

Our covariance matrix estimator is similar in construction to the POET estimator by Fan, Liao, and Mincheva (2013) for discrete time series, except that we block-diagonalize Γ instead of using soft- or hard- thresholding. The latter approach would inevitably introduce additional tuning parameters, which we try to avoid. The same principle is adopted by Fan, Furger, and Xiu (2015) in a similar setting with observable factors.

Equivalently, we can also motivate our estimator from least-square estimation in Stock and Watson (2002), Bai and Ng (2013), and Fan, Liao, and Mincheva (2013). Our estimator can be written as

$$\widehat{\Sigma}^S = t^{-1} F G G^\intercal F^\intercal + \widehat{\Gamma}^S, \quad \widehat{\Gamma} = t^{-1} (\mathcal{Y} - FG)(\mathcal{Y} - FG)^\intercal, \quad \text{and} \quad \widehat{\Gamma}^S = (\widehat{\Gamma}_{ij} 1_{(i,j)\in S}), \tag{5}$$

where $\mathcal{Y} = (\Delta_1^n Y, \Delta_2^n Y, \ldots, \Delta_n^n Y)$ is a $d \times n$ matrix, $G = (g_1, g_2, \ldots, g_n)$ is $\widehat{r} \times n$, $F = (f_1, f_2, \ldots, f_d)^\intercal$ is $d \times \widehat{r}$, and F and G solve the least-squre problem:

$$(F, G) = \arg \min_{f_k, g_i \in \mathbb{R}_{\widehat{r}}} \sum_{i=1}^{n} \sum_{k=1}^{d} \left(\Delta_i^n Y_k - f_k^\intercal g_i\right)^2 = \arg \min_{F \in \mathcal{M}_{d\times\widehat{r}}, G \in \mathcal{M}_{\widehat{r}\times n}} \|\mathcal{Y} - FG\|_{\mathrm{F}}^2$$

subject to

$$d^{-1} F^\intercal F = \mathrm{I}_{\widehat{r}}, \quad GG^\intercal \text{ is an } \widehat{r} \times \widehat{r} \text{ diagonal matrix.}$$

Our least-square estimator is similar to those by Bai and Ng (2002), Bai (2003), and Fan, Liao, and Mincheva (2013), except that we apply the PCA to the $d \times d$ matrix $\mathcal{Y}\mathcal{Y}^\intercal$ instead of the $n \times n$ matrix $\mathcal{Y}^\intercal\mathcal{Y}$. This is mainly because our spot covariance matrices e_s and c_s are time-varying, so that the $n \times n$ matrix is conceptually more difficult to analyze. It is straightforward to verify that $F = d^{1/2}\left(\widehat{\xi}_1, \widehat{\xi}_2, \ldots, \widehat{\xi}_{\widehat{r}}\right)$ and $G = d^{-1} F^\intercal \mathcal{Y}$ are the solutions to this optimization problem, and the estimator given by (5) is the same as that given by (3) and (4).

To determine the number of factors, we propose the following estimator using a penalty function:

$$\widehat{r} = \arg \min_{1 \le j \le r_{\max}} \left(d^{-1} \lambda_j(\widehat{\Sigma}) + j \times g(n, d)\right) - 1,$$

[1] Aït-Sahalia and Xiu (2014) discuss the more general setting where eigenvalues are potentially repeated.

where $r_{\max}$ is some upper bound of $r+1$. The penalty function $g(n,d)$ satisfies two criterions. On the one hand, the penalty cannot dominate the signal, i.e., the value of $d^{-1}\lambda_j(\Sigma)$, when $1 \leq j \leq r$. Since $d^{-1}\lambda_r(\Sigma)$ is $O_p(1)$ as d increases, the penalty should shrink to 0. On the other hand, the penalty should dominate the estimation error as well as $d^{-1}\lambda_{r+1}(\Sigma)$ when $r+1 \leq j \leq d$ to avoid overshooting. The choice of $r_{\max}$ does not play any role in theory. It is only used to avoid an economically nonsensical choice of r in finite sample or in practice.

Our estimator is similar in spirit to that introduced in Bai and Ng (2002). They suggest to estimate r by minimizing the penalized objective function:

$$\widehat{r} = \arg \min_{1 \leq j \leq r_{\max}} (d \times t)^{-1} \left\| \mathcal{Y} - F(j)G(j) \right\|_{\mathrm{F}}^2 + \text{penalty},$$

where the dependence of F and G on j is highlighted. It turns out, perhaps not surprisingly, that

$$(d \times t)^{-1} \left\| \mathcal{Y} - F(j)G(j) \right\|_{\mathrm{F}}^2 = d^{-1} \sum_{i=j+1}^{d} \lambda_i(\widehat{\Sigma}),$$

which is closely related to our proposed objective function. It is, however, easier to use our proposal as it does not involve estimating the sum of many eigenvalues. The proof is also simpler.

There are many alternative methods to determine the number of factors, including Hallin and Liška (2007), Amengual and Watson (2007), Alessi, Barigozzi, and Capasso (2010), Kapetanios (2010), and Onatski (2010). Ahn and Horenstein (2013) propose an estimator by maximizing the ratios of adjacent eigenvalues. Their approach is convenient and it does not involve any penalty function. Unfortunately, the consistency of their estimator requires the random matrix theory established by, e.g., Bai and Yin (1993), so as to establish a sharp convergence rate for the eigenvalue ratio of the sample covariance matrix. Such a theory is not available for semimartingales to the best of our knowledge, we thereby propose the alternative estimator, for which we can establish the desired consistency without using the random matrix theory.

3.3 Asymptotic Theory

Our theory is based on the dual in-fill and diverging dimensionality asymptotics with the number of factors being finite. That is, $\Delta_n \to 0$, $d \to \infty$, and r is fixed but unknown. We first establish the consistency of $\widehat{r}$.

Theorem 2. *Under Assumptions 1 - 4, and suppose that* $d^{-1}m_d = o(1)$, $\Delta_n \log d = o(1)$, $g(n,d) \to 0$, *and* $g(n,d)\left((\Delta_n \log d)^{1/2} + d^{-1}m_d\right)^{-1} \to \infty$, *we have* $\mathbb{P}(\widehat{r} = r) \to 1$.

A choice of the penalty function could be

$$g(n,d) = \mu \left((n^{-1} \log d)^{1/2} + d^{-1} m_d \right)^{\kappa},$$

where μ and κ are some constants and $0 < \kappa < 1$. While it may be difficult to choose these tuning parameters in practice, the covariance matrix estimates are not sensitive to the numbers of factors.

Also, the scree plot offers the rule-of-thumb guide to set r. Practically speaking, r is no different from a "tuning parameter." And it is much easier to interpret r than μ and κ above. In the later portfolio allocation study, we choose a range of rs to compare our covariance matrix estimator with that using observable factors. As long as r is larger than 3, the results do not change much and the interpretation remains the same.

The next theorem establishes the desired consistency of the covariance matrix estimator.

Theorem 3. *Suppose Assumptions 1 - 4 hold. Also, $d^{-1/2}m_d = o(1)$ and $\Delta_n \log d = o(1)$ are satisfied. Suppose $\widehat{r} \to r$ with probability approaching 1, then we have*

$$\left\|\widehat{\Gamma}^S - \Gamma\right\|_{\text{MAX}} = O_p\left((\Delta_n \log d)^{1/2} + d^{-1/2}m_d\right).$$

Moreover, we have

$$\left\|\widehat{\Sigma}^S - \Sigma\right\|_{\text{MAX}} = O_p\left((\Delta_n \log d)^{1/2} + d^{-1/2}m_d\right).$$

Compared to the rate of the regression based estimator in Fan, Furger, and Xiu (2015), i.e., $O_p((\Delta_n \log d)^{1/2})$, the convergence rate of the PCA estimator depends on a new term $d^{-1/2}m_d$, due to the presence of unobservable factors, as can be seen from Theorem 1. We consider the consistency under the entry-wise norm instead of the operator norm, partially because the eigenvalues of Σ themselves grow at the rate of $O(d)$, so that their estimation errors do not shrink to 0, when the dimension d increases exponentially, relative to the sampling frequency Δ_n.

In terms of the portfolio allocation, the precision matrix perhaps plays a more important role than the covariance matrix. For instance, the popular minimum variance portfolio is determined by the inverse of the Σ instead of the Σ itself. Our estimator is not only positive-definite, but is also well-conditioned. This is because the minimum eigenvalue of the estimator is bounded from below with probability approaching 1. The next theorem describes the asymptotic behavior of the precision matrix estimation under the operator norm.

Theorem 4. *Suppose Assumptions 1 - 4 hold. Suppose $d^{-1/2}m_d = o(1)$, $\Delta_n \log d = o(1)$, and $\widehat{r} \to r$ with probability approaching 1, then we have*

$$\left\|\widehat{\Gamma}^S - \Gamma\right\| = O_p\left(m_d(\Delta_n \log d)^{1/2} + d^{-1/2}m_d^2\right).$$

If in addition, $d^{-1/2}m_d^2 = o(1)$ and $m_d(\Delta_n \log d)^{1/2} = o(1)$, then $\lambda_{\min}(\widehat{\Sigma}^S)$ is bounded away from 0 with probability approaching 1, and

$$\left\|(\widehat{\Sigma}^S)^{-1} - \Sigma^{-1}\right\| = O\left(m_d^3\left((\Delta_n \log d)^{1/2} + d^{-1/2}m_d\right)\right).$$

The convergence rate of the regression based estimator in Fan, Furger, and Xiu (2015) with observable factors is $O_p(m_d(\Delta_n \log d)^{1/2})$. In their paper, the eigenvalues of Γ is bounded from above, whereas we relax this assumption in this paper, which explains the extra powers of m_d here. As above, $d^{-1/2}m_d$ reflects the loss due to ignorance of the latent factors.

As a by-product, we can also establish the consistency of factors and loadings up to some matrix transformation:

Theorem 5. *Suppose Assumptions 1 - 4 hold. Suppose* $d^{-1/2}m_d = o(1)$, $\Delta_n \log d = o(1)$, *and* $\widehat{r} \to r$ *with probability approaching 1, then there exists a* $r \times r$ *matrix* H, *such that with probability approaching 1,* H *is invertible,* $\|HH^\intercal - \mathrm{I}_r\| = \|H^\intercal H - \mathrm{I}_r\| = o_p(1)$, *and more importantly,*

$$\|F - \beta H\|_{\mathrm{MAX}} = O_p\left((\Delta_n \log d)^{1/2} + d^{-1/2}m_d\right), \quad \left\|G - H^{-1}\mathcal{X}\right\| = O_p\left((\Delta_n \log d)^{1/2} + d^{-1/2}m_d\right).$$

The presence of the H matrix is due to the indeterminacy of a factor model. Bai and Ng (2013) further impose strong assumptions so as to identify the factors. For instance, one set of identification assumptions suggest that the first few observed asset returns are essentially noisy observations of the factors themselves. For the purpose of covariance matrix estimation, such assumptions are not needed.

4 Monte Carlo Simulations

In the previous section, we have established the theoretical asymptotic results in an ideal setting without market microstructure concerns. This setting is realistic and relevant in practice only for returns sampled at a sufficiently low frequency. We choose this setup mainly to demonstrate the effect of an increasing dimensionality. In this section, we also examine the effect of subsampling on the performance of our estimators in the presence of both asynchronous observations and microstructure noise.

Following the setup in Fan, Furger, and Xiu (2015), we sample 100 paths from a continuous-time r-factor model of d assets specified as:

$$dY_{i,t} = \sum_{j=1}^{r} \beta_{i,j} dX_{j,t} + dZ_{i,t}, \quad dX_{j,t} = b_j dt + \sigma_{j,t} dW_{j,t}, \quad dZ_{i,t} = \gamma_i^\intercal dB_{i,t},$$

where W_j is a standard Brownian motion and B_i is a d-dimensional Brownian motion, for $i = 1, 2, \ldots, d$, and $j = 1, 2, \ldots, r$. They are mutually independent. X_j is the jth unobservable factor. We deem one of the Xs as the market factor, so that its associated βs are positive. The covariance matrix of Z is a block diagonal matrix, denoted by Γ, that is, $\Gamma_{il} = \gamma_i^\intercal \gamma_l$. We allow for time-varying $\sigma_{j,t}$ which evolves according to the following system of equations:

$$d\sigma_{j,t}^2 = \kappa_j(\theta_j - \sigma_{j,t}^2)dt + \eta_j \sigma_{j,t} d\widetilde{W}_{j,t}, \quad j = 1, 2, \ldots, r,$$

where $\widetilde{W}_j$ is a standard Brownian motion with $\mathbb{E}[dW_{j,t} d\widetilde{W}_{j,t}] = \rho_j dt$. We choose $d = 500$ and $r = 3$. In addition, $\kappa = (3,\ 4,\ 5)$, $\theta = (0.05,\ 0.04,\ 0.03)$, $\eta = (0.3,\ 0.4,\ 0.3)$, $\rho = (\text{-}0.60,\ \text{-}0.40,\ \text{-}0.25)$, and $b = (0.05, 0.03, 0.02)$. In the cross-section, we sample $\beta_1 \sim \mathcal{U}[0.25, 1.75]$, and sample $\beta_2, \beta_3 \sim \mathcal{N}(0, 0.5^2)$. The variances on the diagonal of Γ are uniformly generated from $[0.05, 0.20]$, with constant within-block correlations sampled from $\mathcal{U}[0.10, 0.50]$ for each block. In total, there are 20 blocks (of size 25×25) on the diagonal of the residual covariance matrix.

To mimic the effect of microstructure noise and asynchronicity, we add a Gaussian noise with mean zero and variance 0.001^2 to the simulated log prices before censoring. The data are then

censored using Poisson sampling, where the number of observations for each asset is drawn from a truncated log-normal distribution. The parameters of the distribution are calibrated such that its cross-sectional distribution matches the empirical pattern shown in Figure 1 of Lunde, Shephard, and Sheppard (2014).

Table 1 provides the averages of $\|\widehat{\Sigma}-\Sigma\|_{\text{MAX}}$ and $\|(\widehat{\Sigma}^S)^{-1}-\Sigma^{-1}\|$ in various scenarios. We apply the PCA and the regression estimators to the ideal dataset without any noise or asynchronicity. The results are shown in Columns PCA* and REG*. Columns PCA and REG contain the estimation results using the polluted data. In the last column, we report the estimated number of factors with the polluted data. We choose the tuning parameters as $\kappa = 0.5$, $r_{\max} = 20$, and $\mu = 0.04 \times \lambda_{d/2}(\widehat{\Sigma})$. The use of the median eigenvalue $\lambda_{d/2}(\widehat{\Sigma})$ helps adjust the level of average eigenvalues for better accuracy.

We summarize the findings here. First, the values of $\|\widehat{\Sigma} - \Sigma\|_{\text{MAX}}$ in Columns REG and PCA are identical. This is due to the fact that the largest entry-wise errors are likely achieved along the diagonals, and that the estimates on the diagonal are identical to the sample covariance estimates, regardless of whether the factors are observable or not. As to the prevision matrix under the operator norm, i.e., $\|(\widehat{\Sigma}^S)^{-1} - \Sigma^{-1}\|$, the differences between the two estimators are noticeable despite being very small. While the PCA approach uses less information, it can perform equally well as the REG approach. That said, the benefit of using observable factors is apparent from the comparison between Columns REG* and PCA*, as the results based on the PCA* are slightly worse. Secondly, the well-known microstructure effect clearly ruins the estimates when the sampling frequency is as high as every few seconds. Subsampling indeed mitigates the microstructure concerns, while it also raises another concern with a relatively increasing dimensionality – the ratio of cross-sectional dimension against number of observations. The sweet spot of the trade-off appears to be in the range between 15 and 30 minutes. Finally, the number of factors is precisely estimated for most frequencies. Not surprisingly, at both ends of the sampling frequency, the estimates are off.

5 Empirical Work

5.1 Data

We collect from the TAQ database intraday observations of the S&P 500 index constituents from January 2004 to December 2012. The constituents have been changing from time to time. As a result, there are in total 736 stocks. We follow the usual procedure, see, e.g., Aït-Sahalia and Jacod (2014), to clean the data and subsample returns of each asset every 15 minutes. The overnight returns are excluded to avoid dividend issuances and stock splits.

In addition, we collect the Global Industrial Classification Standard (GICS) codes from the Compustat database. These 8-digit codes are assigned to each company in the S&P 500. The code is split into 4 groups of 2 digits. Digits 1-2 describe the company's sector; digits 3-4 describe the industry group; digits 5-6 describe the industry; digits 7-8 describe the sub-industry. The GICS codes

	$\|\widehat{\Sigma}-\Sigma\|_{\text{MAX}}$				$\|(\widehat{\Sigma}^S)^{-1}-\Sigma^{-1}\|$				# of Factors
Freq	REG*	PCA*	REG	PCA	REG*	PCA*	REG	PCA	
5	0.005	0.009	2.371	2.371	0.590	3.242	33.426	33.417	1
15	0.008	0.011	0.806	0.806	0.981	3.211	32.875	32.854	1
30	0.011	0.014	0.414	0.414	1.499	3.408	32.012	31.963	3
60	0.018	0.019	0.221	0.221	2.257	3.478	30.466	30.350	3
300	0.037	0.037	0.075	0.075	7.543	7.429	22.371	22.035	3
900	0.049	0.050	0.061	0.061	13.677	13.486	14.237	14.124	3
1800	0.071	0.071	0.072	0.073	20.850	20.450	14.126	14.456	3
3900	0.108	0.108	0.112	0.112	38.693	40.498	35.884	36.537	20
4680	0.142	0.142	0.150	0.150	52.045	54.002	51.443	51.225	20
11700	0.201	0.201	0.205	0.205	634.357	586.789	511.255	486.668	20

Table 1: Simulation Results

Note: In this table, we report the values of $\|\widehat{\Sigma}-\Sigma\|_{\text{MAX}}$ and $\|(\widehat{\Sigma}^S)^{-1}-\Sigma^{-1}\|$ for each subsampling frequency ranging from one observation every 5 seconds to 2 observations per day. The first column displays the sampling frequencies in seconds. Columns REG* and PCA* report the results of regression and the PCA methods respectively, using synchronous observations without microstructure noise. Columns REG and the PCA are based on the polluted data. Columns REG*, REG, and PCA* all assume 3 factors. The results in the PCA column are obtained by estimating the number of factors first. The last column reports the median estimates of the number of factors.

are used to sort stocks and form blocks of the residual covariance matrices. The GICS codes also change over time. The time series median of the largest block size is 77 for sector-based classification, 38 for industry group, 24 for industry, and 14 for sub-industry categories.

For comparison purpose, we also make use of the observable factors constructed from high-frequency returns, including the market portfolio, the small-minus-big market capitalization (SMB) portfolio, and high-minus-low price-earnings ratio (HML) portfolio in the Fama-French 3 factor model, as well as the daily-rebalanced momentum portfolio formed by sorting stock returns between the past 250 days and 21 days. We follow the same procedure to construct these factors as described on Kenneth French's webpage, see Aït-Sahalia, Kalnina, and Xiu (2014) for more details. We also collect from TAQ 9 industry SDPR ETFs. They are Energy (XLE), Materials (XLB), Industrials (XLI), Consumer Discretionary (XLY), Consumer Staples (XLP), Health Care (XLV), Financial (XLF), Information Technology (XLK), and Utilities (XLU).

5.2 The Number of Factors

Prior to estimating the number of factors, we demonstrate the sparsity and block-diagonal pattern of the residual covariance matrix using various combinations of factors, which server as a rule-of-thumb guide. In Figures 1 and 2, we mark the economically significant entries of the residual covariance estimates for the year 2012, after removing the part driven by 1, 4, 10, and 13 PCA-based factors, respectively. The criterion of the economic significance is that the correlation is at least 0.15 for at

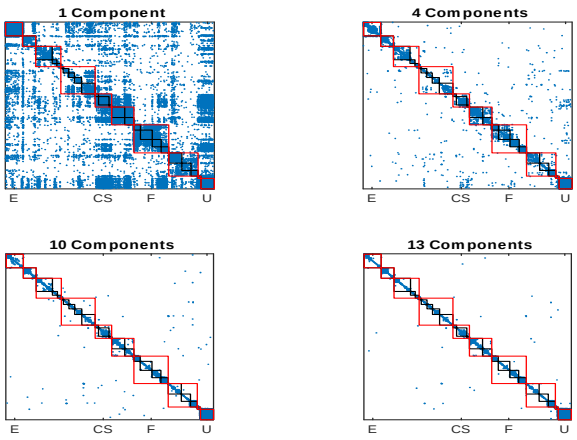

Figure 1: The Sparsity Pattern of the Residual Covariance Matrices

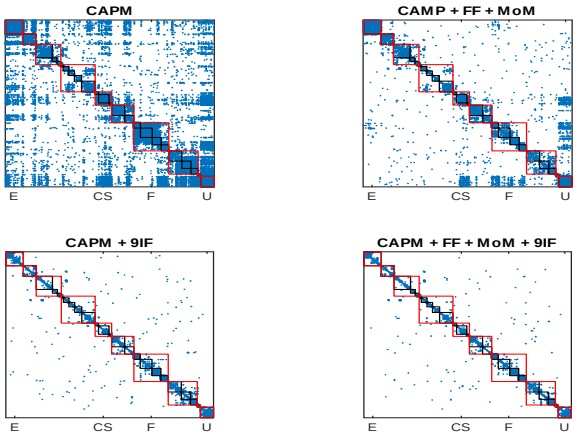

Figure 2: The Sparsity Pattern of the Residual Covariance Matrices

least 1/3 of the year. These two thresholds and the choice of the year 2012 are entirely arbitrary. Varying these numbers do not change the pattern and the message of the plots. We also compare these plots with those based on observable factors. More specifically, our benchmark 1-factor model is the CAPM. For the 4-factor model, we use 3 Fama-French portfolios plus the momentum portfolio. The 10-factor model is based on 1 market portfolio and 9 industrial ETFs. The 13-factor model uses all factors.

Note: The figure displays the significant entries of the residual covariance matrices, relative to 1, 4, 10, and 13 latent factors. The red (resp. black) squares highlight those stocks that belong to the same sector (resp. industry group).

Note: The figure displays the significant entries of the residual covariance matrices, relative to 1, 4, 10, and 13 observable factors. The red (resp. black) squares highlight those stocks that belong to the same

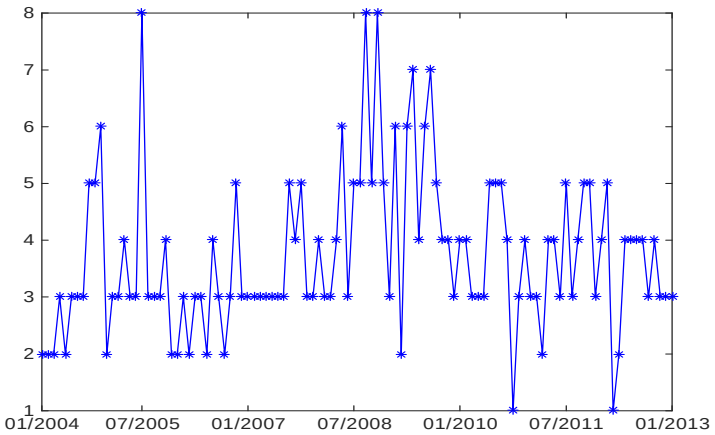

Figure 3: Estimates of the Number of Factors

sector (resp. industry group). CAPM denotes one factor case using the market portfolio, FF refers to the two additional Fama-French factors, MoM denotes the momentum factor, whereas the 9IF refers to the 9 industrial ETF factors. They are Energy (XLE), Materials (XLB), Industrials (XLI), Consumer Discretionary (XLY), Consumer Staples (XLP), Health Care (XLV), Financial (XLF), Information Technology (XLK), and Utilities (XLU).

We find that the PCA approach is very effective in identifying the latent factors. The residual covariance matrix demonstrates a clear block-diagonal pattern after removing as few as 4 latent factors. The residual correlations are likely due to idiosyncrasies within sectors or industrial groups. This pattern verifies the low-rank plus sparsity structure we impose. Instead of thresholding all off-diagonal entries as suggested by the strict factor model, our proposal maintains within-sector or within-industry correlations, hence produces more accurate estimates. As documented in Fan, Furger, and Xiu (2015), there is a similar pattern with observable factors, but apparently, more of such factors are necessary to obtain the same degree of the sparsity obtained by the PCA approach.

We then use our estimator to determine the number of factors for each month. The time series plot is shown in Figure 3. The times series is relatively stable, pointing out 2 - 5 factors for most of the sample periods. The result echoes with the scree plot shown in Aït-Sahalia and Xiu (2014). This finding also agrees with the pattern in the residual sparsity plot.

Note: This figure plots the time series of the estimated number of factors using the PCA. The tuning parameters in the penalty function are $\mu = 0.04 \times \lambda_{d/2}(\widehat{\Sigma})$, $\kappa = 0.5$, and $r_{\max} = 20$.

5.3 In-Sample R^2 Comparison

We now compare the variation explained by an increasing number of latent factors with the variation explained by the same number of observable factors. We calculate their in-sample R^2 respectively for each stock and for each month, and plot the time series of their cross-sectional medians in Figure

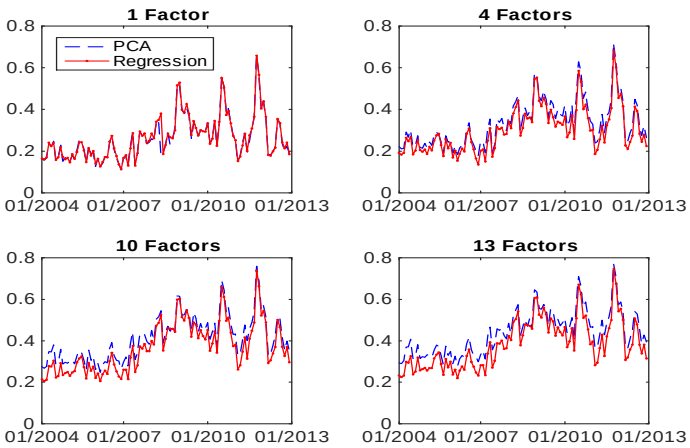

Figure 4: In-Sample R^2 Comparison

4. Not surprisingly, the first latent factor agrees with the market portfolio. Despite that it is not a portfolio, it explains as much variation as the market portfolio. Nevertheless, when more factors are included, both the latent factors and the observable factors can explain more variation, with the former explaining slightly more. This in fact suggests that the observable factors are rather effective in capturing the latent common factors.

Note: This figure plots the time series of the cross-sectional medians of R^2s based on the latent factors identified from the PCA, as well as those based on the observable factors. The number of factors refers to the number of latent components from the PCA approach and the number of portfolios used in the regression approach.

5.4 Out-of-Sample Portfolio Allocation Study

In this section, we bring together our covariance estimates to an empirical test-drive. We consider the following constrained portfolio allocation exercise:

$$\min_{w} w^{\intercal}\widehat{\Sigma}^{S}w, \quad \text{subject to } \omega^{\intercal}1 = 1, \|\omega\|_1 \leq \gamma, \tag{6}$$

where $\|\omega\|_1 \leq \gamma$ imposes an exposure constraint, see, e.g., Jagannathan and Ma (2003) and Fan, Zhang, and Yu (2012). When $\gamma = 1$, the optimal portfolio allows no short-sales, i.e., all portfolio weights are non-negative. When γ is small and binding, the optimal portfolio is sparse, i.e., many weights are zero. When γ is no longer binding, the optimal portfolio coincides with the global minimum variance portfolio.

For each month from February 2004 to December 2012, we build our portfolio based on the covariance estimates in the past month. This amounts to assuming that $\widehat{\Sigma}_t^S \approx \mathrm{E}_t(\Sigma_{t+1})$, which is a common strategy in practice. We compare the out-of-sample performance of the portfolio allocation problem (6) with a range of exposure constraints. The results are shown in Figure 5.

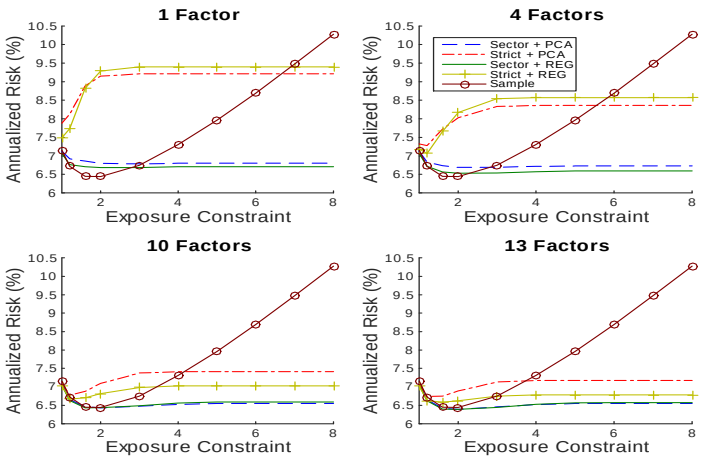

Figure 5: Out-of-Sample Risk of the Portfolio

Note: This figure compares the time series average of the out-of-sample monthly volatility from 2004 and 2012. The x-axis is the exposure constraint γ in the optimization problem (6). The results are based on 5 covariance matrix estimators, including the sample covariance matrix (Sample), the PCA approach with sector-grouped block-diagonal residual covariance (Sector + PCA), PCA with diagonal residual covariance (Strict + PCA), and their regression counterparts (Sector + REG, Strict + REG). The number of factors refers to the number of principal components for the PCA approach and the number of portfolios factors for the regression approach.

We find that for the purpose of portfolio allocation, the PCA approach performs as well as the regression method. Their performance further improves when combined with sector-based block-diagonal structure of the residual covariance matrix. The sample covariance matrix based allocation only performs reasonably well when the exposure constraint is very tight. As the constraint relaxes, more stocks are selected into the portfolio, so that the in-sample risk of the portfolio decreases. However, the risk of the sample covariance based portfolio increases out-of-sample, suggesting that the covariance matrix estimates are ill-conditioned that the allocation becomes noisy and unstable. Both the PCA and the regression approach produce stable out-of-sample risk, as the exposure constraint relaxes. We also build up an equal-weight portfolio, and its annualized risk is 17.89%. We did not plot it as it is independent of the exposure constraints and the numbers of factors. taowangchen2013

6 Conclusion

We propose a simple PCA based estimator of the large covariance matrix using high frequency returns. Our model is semiparametric, allowing latent factors with arbitrary heteroscedasticity. The estimator is positive-definite and well-conditioned. We also provide a new estimator of the number of latent factors. We justify the consistency of these estimators using the dual in-fill and diverging dimensionality asymptotics, which sheds light on both the curse and blessings of the dimensionality.

Empirically, we document a latent low-rank and sparsity structure behind the covariances of the asset returns. Our comparison with observable factors also suggest that the Fama-French factors, the momentum factor, and the industrial portfolios together, approximate the span of the latent factors very well.

References

Ahn, S. C., and A. R. Horenstein (2013): "Eigenvalue Ratio Test for the Number of Factors," *Econometrica*, 81(3), 1203–1227.

Aït-Sahalia, Y., J. Fan, and D. Xiu (2010): "High-Frequency Covariance Estimates with Noisy and Asynchronous Data," *Journal of the American Statistical Association*, 105, 1504–1517.

Aït-Sahalia, Y., and J. Jacod (2014): *High Frequency Financial Econometrics*. Princeton University Press.

Aït-Sahalia, Y., I. Kalnina, and D. Xiu (2014): "The Idiosyncratic Volatility Puzzle: A Reassessment at High Frequency," Discussion paper, The University of Chicago.

Aït-Sahalia, Y., and D. Xiu (2014): "Principal Component Analysis of High Frequency Data," Discussion paper, Princeton University and the University of Chicago.

Alessi, L., M. Barigozzi, and M. Capasso (2010): "Improved Penalization for Determining the Number of Factors in Approximate Factor Models," *Statistics and Probability Letters*, 80, 1806–1813.

Amengual, D., and M. W. Watson (2007): "Consistent Estimation of the Number of Dynamic Factors in a Large N and T Panel," *Journal of Business & Economic Statistics*, 25(1), 91–96.

Bai, J. (2003): "Inferential Theory for Factor models of Large Dimensions," *Econometrica*, 71, 135–171.

Bai, J., and S. Ng (2002): "Determining the Number of Factors in Approximate Factor Models," *Econometrica*, 70, 191–221.

——— (2013): "Principal components estimation and identification of static factors," *Journal of Econometrics*, 176(1), 18–29.

Bai, Z. D., and Y. Q. Yin (1993): "Limit of the Smallest Eigenvalue of a Large Dimensional Sample Covariance Matrix," *The Annals of Probability*, 21(3), 1275–1294.

Barndorff-Nielsen, O. E., P. R. Hansen, A. Lunde, and N. Shephard (2011): "Multivariate Realised Kernels: Consistent Positive Semi-Definite Estimators of the Covariation of Equity Prices with Noise and Non-Synchronous Trading," *Journal of Econometrics*, 162, 149–169.

Barndorff-Nielsen, O. E., and N. Shephard (2004): "Econometric Analysis of Realized Covariation: High Frequency Based Covariance, Regression, and Correlation in Financial Economics," *Econometrica*, 72(3), 885–925.

Bibinger, M., N. Hautsch, P. Malec, and M. Reiss (2014): "Estimating the quadratic covariation matrix from noisy observations: Local method of moments and efficiency," *The Annals of Statistics*, 42(4), 1312 – 1346.

Bickel, P. J., and E. Levina (2008a): "Covariance Regularization by Thresholding," *Annals of Statistics*, 36(6), 2577–2604.

——— (2008b): "Regularized Estimation of Large Covariance Matrices," *Annals of Statistics*, 36, 199–227.

Cai, T., and W. Liu (2011): "Adaptive Thresholding for Sparse Covariance Matrix Estimation," *Journal of the American Statistical Association*, 106, 672–684.

Chamberlain, G., and M. Rothschild (1983): "Arbitrage, Factor Structure, and Mean-Variance Analysis on Large Asset Markets," *Econometrica*, 51, 1281–1304.

Chen, N.-F., R. Roll, and S. A. Ross (1986): "Economic forces and the stock market," *Journal of Business*, 50(1).

Christensen, K., S. Kinnebrock, and M. Podolskij (2010): "Pre-averaging estimators of the ex-post covariance matrix in noisy diffusion models with non-synchronous data," *Journal of Econometrics*, 159, 116–133.

Connor, G., and R. Korajczyk (1988): "Risk and Return in an Equilibrium APT: Application of a New Test Methodology," *Journal of Financial Economics*, 21, 255–289.

Davis, C., and W. M. Kahan (1970): "The rotation of eigenvectors by a perturbation. III.," *SIAM Journal on Numerical Analysis*, 7, 1–46.

Fama, E. F., and K. R. French (1993): "Common Risk Factors in the Returns on Stocks and Bonds," *Journal of Financial Economics*, 33, 3–56.

Fan, J., Y. Fan, and J. Lv (2008): "High Dimensional Covariance Matrix Estimation using a Factor Model," *Journal of Econometrics*, 147, 186–197.

Fan, J., A. Furger, and D. Xiu (2015): "Incorporating Global Industrial Classification Standard into Portfolio Allocation: A Simple Factor-Based Large Covariance Matrix Estimator with High Frequency Data," *Journal of Business and Economic Statistics, forthcoming.*

Fan, J., Y. Liao, and M. Mincheva (2011): "High-Dimensional Covariance Matrix Estimation in Approximate Factor Models," *Annals of Statistics*, 39(6), 3320–3356.

——— (2013): "Large Covariance Estimation by Thresholding Principal Orthogonal Complements," *Journal of the Royal Statistical Society, B*, 75, 603–680.

Fan, J., Y. Liao, and W. Wang (2014): "Projected Principal Component Analysis in Factor Models," Discussion paper, Princeton University.

Fan, J., and W. Wang (2014): "Asymptotics of Empirical Eigenstructure for Ultra-high Dimensional Spiked Covariance Model," Discussion paper, Princeton University.

Fan, J., J. Zhang, and K. Yu (2012): "Vast portfolio selection with gross-exposure constraints," *Journal of the American Statistical Association*, 107, 592–606.

Forni, M., M. Hallin, M. Lippi, and L. Reichlin (2000): "The Generalized Dynamic-Factor Model: Identification and Estimation," *The Review of Economics and Statistics*, 82, 540–554.

——— (2004): "The Generalized Dynamic Factor Model: Consistency and Rates," *Journal of Econometrics*, 119(2), 231–255.

Forni, M., and M. Lippi (2001): "The Generalized Dynamic Factor Model: Representation Theory," *Econometric Theory*, 17, 1113–1141.

Hallin, M., and R. Liška (2007): "Determining the Number of Factors in the General Dynamic Factor Model," *Journal of the American Statistical Association*, 102(478), 603–617.

Horn, R. A., and C. R. Johnson (2013): *Matrix Analysis*. Cambridge University Press, second edn.

Jacod, J., and P. Protter (2011): *Discretization of Processes*. Springer-Verlag.

Jagannathan, R., and T. Ma (2003): "Risk Reduction in Large Portfolios: Why Imposing the Wrong Constraints Helps," *The Journal of Finance*, 58, 1651–1684.

Kapetanios, G. (2010): "A Testing Procedure for Determining the Number of Factors in Approximate Factor Models," *Journal of Business & Economic Statistics*, 28, 397–409.

Ledoit, O., and M. Wolf (2004a): "Honey, I shrunk the sample covariance matrix," *Journal of Portfolio Management*, 30, 110–119.

——— (2004b): "A well-conditioned estimator for large-dimensional covariance matrices," *Journal of Multivariate Analysis*, 88, 365–411.

——— (2012): "Nonlinear shrinkage estimation of large-dimensional covariance matrices," *The Annals of Statistics*, 40, 1024–1060.

Lunde, A., N. Shephard, and K. Sheppard (2014): "Econometric Analysis of Vast Covariance Matrices Using Composite Realized Kernels," Discussion paper, Aarhus University.

Merton, R. C. (1973): "An Intertemporal Capital Asset Pricing Model," *Econometrica*, 41, 867–887.

Mykland, P. A., and L. Zhang (2006): "ANOVA for Diffusions and Itô Processes," *Annals of Statistics*, 34, 1931–1963.

Onatski, A. (2010): "Determining the Number of Factors from Empirical Distribution of Eigenvalues," *Review of Economics and Statistics*, 92, 1004–1016.

Pelger, M. (2015a): "Large-dimensional factor modeling based on high-frequency observations," Discussion paper, Stanford University.

——— (2015b): "Understanding Systematic Risk: A High-Frequency Approach," Discussion paper, Stanford University.

Reiss M., V. Todorov, and G. E. Tauchen (2015): "Nonparametric Test for a Constant Beta between Itô Semi-martingales based on High-Frequency Data," *Stochastic Processes and their Applications, forthcoming*.

Ross, S. A. (1976): "The Arbitrage Theory of Capital Asset Pricing," *Journal of Economic Theory*, 13, 341–360.

Shephard, N., and D. Xiu (2012): "Econometric analysis of multivariate realized QML: Estimation of the covariation of equity prices under asynchronous trading," Discussion paper, University of Oxford and University of Chicago.

Stock, J. H., and M. W. Watson (2002): "Forecasting using Principal Components from a Large Number of Predictors," *Journal of American Statistical Association*, 97, 1167–1179.

Tao, M., Y. Wang, and X. Chen (2013): "Fast Convergence Rates in Estimating Large Volatility Matrices Using High-Frequency Financial Data," *Econometric Theory*, 29(4), 838–856.

Tao, M., Y. Wang, Q. Yao, and J. Zou (2011): "Large Volatility Matrix Inference via Combining Low-Frequency and High-Frequency Approaches," *Journal of the American Statistical Association*, 106, 1025–1040.

Tao, M., Y. Wang, and H. H. Zhou (2013): "Optimal Sparse Volatility Matrix Estimation for High-Dimensional Itô Processes with Measurement Errors," *Annals of Statistics*, 41(1), 1816–1864.

Todorov, V., and T. Bollerslev (2010): "Jumps and Betas: A New Framework for Disentangling and Estimating Systematic Risks," *Journal of Econometrics*, 157, 220–235.

Zhang, L. (2011): "Estimating Covariation: Epps Effect and Microstructure Noise," *Journal of Econometrics*, 160, 33–47.

Zhou, H. H., T. Cai, and Z. Ren (2014): "Estimating Structured High-Dimensional Covariance and Precision Matrices: Optimal Rates and Adaptive Estimation," Discussion paper, Yale University.

Appendix A Mathematical Proofs

Appendix A.1 Proof of Theorem 1

Proof of Theorem 1. First, we write $\mathrm{B} = \beta\sqrt{\mathrm{E}} = (\mathrm{b}_1, \mathrm{b}_2, \ldots, \mathrm{b}_r)$ with $\|b_j\|$s sorted in a descending order. Note that $\left\{\|\mathrm{b}_j\|^2, 1 \le j \le r\right\}$ are the non-zero eigenvalues of $\mathrm{BB}^\intercal$. Therefore by Weyl's inequalities, we have

$$|\lambda_j(\Sigma) - \|\mathrm{b_j}\|^2| \le \|\Gamma\|, 1 \le j \le r; \quad \text{and} \quad |\lambda_j(\Sigma)| \le \|\Gamma\|, r+1 \le j \le d.$$

On the other hand, the non-zero eigenvalues of $\mathrm{BB}^\intercal$ are the eigenvalues of $\mathrm{B}^\intercal\mathrm{B}$. By Weyl's inequalities and Assumption 4, we have, for $1 \le j \le r$,

$$\left|d^{-1}\lambda_j\left(\mathrm{B}^\intercal\mathrm{B}\right) - \lambda_j(\mathrm{E})\right| = \left|d^{-1}\lambda_j\left(\sqrt{\mathrm{E}}\beta^\intercal\beta\sqrt{\mathrm{E}}\right) - \lambda_j(\mathrm{E})\right| \le \|\mathrm{E}\|\left\|d^{-1}\beta^\intercal\beta - \mathrm{I_r}\right\| = o(1).$$

Therefore, $\|\mathrm{b}_j\|^2 = O(d)$, and $K'd \le \lambda_j(\Sigma) \le Kd$, for $1 \le j \le r$. Since $\|\Gamma\| \le \|\Gamma\|_1 \le Km_d$ and $\lambda_j(\Sigma) \ge \lambda_j(\Gamma)$ for $1 \le j \le d$, it follows that $K' \le \lambda_j(\Sigma) \le Km_d$, for $r+1 \le j \le d$. This implies that $d^{-1}\lambda_j(\Sigma) \ge d^{-1}\lambda_r(\Sigma) \ge K'$, for $1 \le j \le r$; $d^{-1}\lambda_j(\Sigma) \le d^{-1}m_d$, for $r+1 \le j \le d$. Since $d^{-1/2}m_d = o(1)$, it follows that $d^{-1}m_d < d^{-1/2}m_d < K'$. Therefore, we have, as $d \to \infty$:

$$\bar{r} = \arg\min_{1\le j\le d}\left(d^{-1}\lambda_j(\Sigma) + jd^{-1/2}m_d\right) - 1 \to r.$$

Next, by the Sin theta theorem in Davis and Kahan (1970), we have

$$\left\|\xi_j - \frac{\mathrm{b}_j}{\|\mathrm{b}_j\|}\right\| \le \frac{K\|\Gamma\|}{\min\left(\left|\lambda_{j-1}(\Sigma) - \|\mathrm{b_j}\|^2\right|, \left|\lambda_{j+1}(\Sigma) - \|\mathrm{b_j}\|^2\right|\right)}.$$

By the triangle inequality, we have

$$\left|\lambda_{j-1}(\Sigma) - \|\mathrm{b_j}\|^2\right| \ge \left|\|\mathrm{b_{j-1}}\|^2 - \|\mathrm{b_j}\|^2\right| - \left|\lambda_{j-1}(\Sigma) - \|\mathrm{b_{j-1}}\|^2\right| \ge \left|\|\mathrm{b_{j-1}}\|^2 - \|\mathrm{b_j}\|^2\right| - \|\Gamma\| > Kd,$$

because for any $1 \le j \le r$, the proof above shows that $\|\mathrm{b_{j-1}}\|^2 - \|\mathrm{b_j}\|^2 = d\left(\lambda_{j-1}(\mathrm{E}) - \lambda_j(\mathrm{E})\right) + o(1)$. Similarly, $\left|\lambda_{j+1}(\Sigma) - \|\mathrm{b_j}\|^2\right| > Kd$, when $j \le r-1$. When $j = r$, we have $\|\mathrm{b_r}\|^2 - \lambda_{j+1}(\Sigma) \ge \|\mathrm{b_r}\|^2 - \|\Gamma\| > Kd$. Therefore, it implies that

$$\left\|\xi_j - \frac{\mathrm{b}_j}{\|\mathrm{b}_j\|}\right\| = O\left(d^{-1}m_d\right), \quad 1 \le j \le r.$$

This, along with the triangle inequality, $\|\mathrm{B}\|_{\mathrm{MAX}} \le \|\beta\|_{\mathrm{MAX}}\left\|\mathrm{E}^{1/2}\right\|_1 \le K$, and $\|\cdot\|_{\mathrm{MAX}} \le \|\cdot\|$, implies that for $1 \le j \le r$,

$$\left\|\xi_j\right\|_{\mathrm{MAX}} \le \left\|\frac{\mathrm{b}_j}{\|\mathrm{b}_j\|}\right\|_{\mathrm{MAX}} + O\left(d^{-1}m_d\right) \le O(d^{-1/2}) + O\left(d^{-1}m_d\right).$$

Since $\bar{r} = r$, for d sufficiently large, by triangle inequalities and that $\|\cdot\|_{\mathrm{MAX}} \le \|\cdot\|$ again, we have

$$\left\|\sum_{j=1}^r \lambda_j\xi_j\xi_j^\intercal - \mathrm{BB}^\intercal\right\|_{\mathrm{MAX}} \le \sum_{j=1}^r \|b_j\|^2\left\|\frac{\mathrm{b}_j}{\|\mathrm{b}_j\|}\right\|_{\mathrm{MAX}}\left\|\xi_j - \frac{\mathrm{b}_j}{\|\mathrm{b}_j\|}\right\|_{\mathrm{MAX}} + \sum_{j=1}^r |\lambda_j - \|b_j\|^2|\left\|\xi_j\xi_j^\intercal\right\|_{\mathrm{MAX}}$$

$$
\begin{aligned}
&+\sum_{j=1}^{r}\|b_j\|^2\left\|\xi_j\right\|_{\mathrm{MAX}}\left\|\xi_j-\frac{\mathrm{b}_j}{\|\mathrm{b}_j\|}\right\|_{\mathrm{MAX}}\\
\leq& Kd^{-1/2}m_d.
\end{aligned}
$$

Hence, since $\Sigma=\sum_{j=1}^{d}\lambda_j\xi_j\xi_j^{\intercal}$, it follows that

$$
\left\|\sum_{j=r+1}^{d}\lambda_j\xi_j\xi_j^{\intercal}-\Gamma\right\|_{\mathrm{MAX}}\leq Kd^{-1/2}m_d,
$$

which concludes the proof. □

Appendix A.2 Proof of Theorem 2

Throughout the proofs of Theorems 2 to 5, we will impose the assumption that $\|\beta\|_{\mathrm{MAX}}$, $\|\Gamma\|_{\mathrm{MAX}}$, $\|\mathrm{E}\|_{\mathrm{MAX}}$, $\|\mathcal{X}\|_{\mathrm{MAX}}$, $\|\mathcal{Z}\|_{\mathrm{MAX}}$, are bounded by K uniformly across time and dimensions. This is due to Assumption 1, the fact that X and Z are continuous, and the localization argument in Section 4.4.1 of Jacod and Protter (2011).

We need one lemma on the concentration inequalities for continuous Itô semimartingales.

Lemma 1. *Suppose Assumptions 1 and 2 hold, then we have*

$$
(i)\quad \max_{1\leq l,k\leq d}\left|\sum_{i=1}^{[t/\Delta_n]}(\Delta_i^nZ_l)(\Delta_i^nZ_k)-\int_0^t g_{s,lk}ds\right|=O_p\left((\Delta_n\log d)^{1/2}\right), \tag{A.1}
$$

$$
(ii)\quad \max_{1\leq j\leq r,1\leq l\leq d}\left|\sum_{i=1}^{[t/\Delta_n]}(\Delta_i^nX_j)(\Delta_i^nZ_l)\right|=O_p\left((\Delta_n\log d)^{1/2}\right), \tag{A.2}
$$

$$
(iii)\quad \max_{1\leq j\leq r,1\leq l\leq r}\left|\sum_{i=1}^{[t/\Delta_n]}(\Delta_i^nX_j)(\Delta_i^nX_l)-\int_0^t e_{s,jl}ds\right|=O_p\left((\Delta_n\log d)^{1/2}\right). \tag{A.3}
$$

Proof of Lemma 1. The proof of this lemma is given by that of (i), (iii), (iv) of Lemma 2 in Fan, Furger, and Xiu (2015). □

Proof of Theorem 2. We first recall some notation introduced in the main text. Let $n=[t/\Delta_n]$. Suppose that $\mathcal{Y}=(\Delta_1^nY,\Delta_2^nY,\ldots,\Delta_n^nY)$ is a $d\times n$ matrix, where $\Delta_i^nY=Y_{i\Delta_n}-Y_{(i-1)\Delta_n}$. Similarly, $\mathcal{X}$ and $\mathcal{Z}$ are $r\times n$ and $d\times n$ matrices, respectively. Therefore, we have $\mathcal{Y}=\beta\mathcal{X}+\mathcal{Z}$ and $\widehat{\Sigma}=t^{-1}\mathcal{Y}\mathcal{Y}^{\intercal}$. Let $f(j)=d^{-1}\lambda_j(\widehat{\Sigma})+j\times g(n,d)$. Suppose $\mathrm{R}=\{j|1\leq j\leq k_{\max},j\neq r\}$.

Note that using $\|\beta\|\leq d^{1/2}\|\beta\|_{\mathrm{MAX}}=O(d^{1/2})$ and $\|\Gamma\|_{\infty}\leq Km_d$ we have

$$
\begin{aligned}
\|\mathcal{Y}\mathcal{Y}^{\intercal}-\beta\mathcal{X}\mathcal{X}^{\intercal}\beta^{\intercal}\|\leq&\|\mathcal{Z}\mathcal{X}^{\intercal}\beta^{\intercal}\|+\|\beta\mathcal{X}\mathcal{Z}^{\intercal}\|+\|\mathcal{Z}\mathcal{Z}^{\intercal}-\Gamma\|+\|\Gamma\|\\
\leq&Kd^{1/2}\|\beta\|\,\|\mathcal{Z}\mathcal{X}^{\intercal}\|_{\mathrm{MAX}}+d\,\|\mathcal{Z}\mathcal{Z}^{\intercal}-\Gamma\|_{\mathrm{MAX}}+\|\Gamma\|_{\infty}\\
=&O_p\left(d(\Delta_n\log d)^{1/2}+m_d\right).
\end{aligned}
$$

where we use the following bounds, implied by Lemma 1:

$$\|\mathcal{Z}\mathcal{Z}^\intercal - \Gamma\|_{\mathrm{MAX}} = \max_{1\le k,l\le d}\left(\left|\sum_{i=1}^{n}(\Delta_i^n Z_l)(\Delta_i^n Z_k) - \int_0^t g_{s,lk}ds\right|\right) = O_p((\Delta_n \log d)^{1/2}), \text{ and}$$
$$\|\mathcal{Z}\mathcal{X}^\intercal\|_{\mathrm{MAX}} = O_p((\Delta_n \log d)^{1/2}).$$

Therefore, by Weyl's inequality we have for $1 \le j \le r$,

$$|\lambda_j(\widehat{\Sigma}) - \lambda_j(t^{-1}\beta\mathcal{X}\mathcal{X}^\intercal\beta^\intercal)| = O_p\left(d(\Delta_n \log d)^{1/2} + m_d\right).$$

On the other hand, the non-zero eigenvalues of $t^{-1}\beta\mathcal{X}\mathcal{X}^\intercal\beta^\intercal$ are identical to the eigenvalues of $t^{-1}\sqrt{\mathcal{X}\mathcal{X}^\intercal}\beta^\intercal\beta\sqrt{\mathcal{X}\mathcal{X}^\intercal}$. By Weyl's inequality again, we have for $1 \le j \le r$,

$$\left|d^{-1}\lambda_j\left(t^{-1}\sqrt{\mathcal{X}\mathcal{X}^\intercal}\beta^\intercal\beta\sqrt{\mathcal{X}\mathcal{X}^\intercal}\right) - \lambda_j(t^{-1}\mathcal{X}\mathcal{X}^\intercal)\right| \le t^{-1}\|\mathcal{X}\mathcal{X}^\intercal\|\left\|d^{-1}\beta^\intercal\beta - \mathrm{I}_r\right\| = o_p(1),$$

where we use

$$\|\mathcal{X}\| = \sqrt{\lambda_{\max}(\mathcal{X}\mathcal{X}^\intercal)} \le r^{1/2}\max_{1\le l,j\le r}\left|\sum_{i=1}^{n}(\Delta_i^n X_l)(\Delta_i^n X_j)\right|^{1/2} = O_p(1). \tag{A.4}$$

Also, for $1 \le j \le r$, by Weyl's inequality and Lemma 1, we have

$$|\lambda_j(t^{-1}\mathcal{X}\mathcal{X}^\intercal) - \lambda_j(\mathrm{E})| \le \left\|t^{-1}\mathcal{X}\mathcal{X}^\intercal - \mathrm{E}\right\| = O_p\left((\Delta_n \log d)^{1/2}\right).$$

Combining the above inequalities, we have for $1 \le j \le r$,

$$|d^{-1}\lambda_j(\widehat{\Sigma}) - \lambda_j(\mathrm{E})| \le O_p\left((\Delta_n \log d)^{1/2} + d^{-1}m_d\right) + o_p(1).$$

Therefore, for $1 \le j < r$, we have

$$\lambda_{j+1}(\mathrm{E}) - o_p(1) < d^{-1}\lambda_{j+1}(\widehat{\Sigma}) < \lambda_{j+1}(\mathrm{E}) + o_p(1) < \lambda_j(\mathrm{E}) - o_p(1) < d^{-1}\lambda_j(\widehat{\Sigma}). \tag{A.5}$$

Next, note that

$$\mathcal{Y}\mathcal{Y}^\intercal = \tilde{\beta}\mathcal{X}\mathcal{X}^\intercal\tilde{\beta}^\intercal + \mathcal{Z}\left(\mathrm{I}_n - \mathcal{X}^\intercal(\mathcal{X}\mathcal{X}^\intercal)^{-1}\mathcal{X}\right)\mathcal{Z}^\intercal$$

where $\tilde{\beta} = \beta + \mathcal{Z}\mathcal{X}^\intercal(\mathcal{X}\mathcal{X}^\intercal)^{-1}$. Since $\mathrm{rank}(\tilde{\beta}\mathcal{X}\mathcal{X}^\intercal\tilde{\beta}^\intercal) = r$, and by (4.3.2b) of Theorem 4.3.1 and (4.3.14) of Corollary 4.3.12 in Horn and Johnson (2013), we have for $r+1 \le j \le d$,

$$\lambda_j(\mathcal{Y}\mathcal{Y}^\intercal) \le \lambda_{j-r}\left(\mathcal{Z}\left(\mathrm{I}_n - \mathcal{X}^\intercal(\mathcal{X}\mathcal{X}^\intercal)^{-1}\mathcal{X}\right)\mathcal{Z}^\intercal\right) + \lambda_{r+1}(\tilde{\beta}\mathcal{X}\mathcal{X}^\intercal\tilde{\beta}^\intercal) \le \lambda_{j-r}(\mathcal{Z}\mathcal{Z}^\intercal) \le \lambda_1(\mathcal{Z}\mathcal{Z}^\intercal).$$

Since by Lemma 1 we have

$$\begin{aligned}\lambda_1(\mathcal{Z}\mathcal{Z}^\intercal) = \|\mathcal{Z}\mathcal{Z}^\intercal\| \le \|\mathcal{Z}\mathcal{Z}^\intercal\|_\infty &\le \max_{1\le j,l\le d}\{d|(\mathcal{Z}\mathcal{Z}^\intercal - \Gamma)_{jl}| + m_d|\Gamma_{jl}|\} \\ &= O_p(d(\Delta_n \log d)^{1/2} + m_d),\end{aligned} \tag{A.6}$$

it thus implies that for $r+1 \leq j \leq d$, there exists some $K > 0$, such that

$$d^{-1}\lambda_j(\widehat{\Sigma}) \leq K(\Delta_n \log d)^{1/2} + Kd^{-1}m_d.$$

In sum, for $1 \leq j \leq r$,

$$f(j) - f(r+1) = d^{-1}\left(\lambda_j(\widehat{\Sigma}) - \lambda_{r+1}(\widehat{\Sigma})\right) + (j-r-1)g(n,d) > \lambda_j(\mathrm{E}) + o_p(1) > K,$$

for some $K > 0$. Since $g(n,d)\left((\Delta_n \log d)^{1/2} + d^{-1}m_d\right)^{-1} \to \infty$, it follows that for $r+1 < j \leq d$,

$$\mathbb{P}\left(f(j) < f(r+1)\right) = \mathbb{P}\left((j-r-1)g(n,d) < d^{-1}\left(\lambda_{r+1}(\widehat{\Sigma}) - \lambda_j(\widehat{\Sigma})\right)\right) \to 0.$$

This establishes the desired result. □

Appendix A.3 Proof of Theorem 3

First, we can assume $\widehat{r} = r$. Since it holds with probability approaching 1 as established by Theorem 2, a simple conditioning argument, see, e.g., footnote 5 of Bai (2003), is sufficient to show this is without loss of rigor. Recall that

$$\Lambda = \mathrm{Diag}\left(\widehat{\lambda}_1, \widehat{\lambda}_2, \ldots, \widehat{\lambda}_r\right), \quad F = d^{1/2}\left(\widehat{\xi}_1, \widehat{\xi}_2, \ldots, \widehat{\xi}_r\right), \quad \text{and} \quad G = d^{-1}F^\intercal \mathcal{Y}.$$

We write

$$H = t^{-1}\mathcal{X}\mathcal{X}^\intercal \beta^\intercal F \Lambda^{-1}.$$

It is easy to verify that

$$\begin{aligned}
&\widehat{\Sigma}F = F\Lambda, \quad GG^\intercal = td^{-1} \times \Lambda, \quad F^\intercal F = d \times \mathrm{I}_r, \quad \text{and} \\
&\widehat{\Gamma} = t^{-1}\left(\mathcal{Y} - FG\right)\left(\mathcal{Y} - FG\right)^\intercal = t^{-1}\mathcal{Y}\mathcal{Y}^\intercal - d^{-1}F\Lambda F^\intercal.
\end{aligned}$$

We now need a few more lemmas.

Lemma 2. *Under Assumptions 1 - 4, $d^{-1/2}m_d = o(1)$, and $\Delta_n \log d = o(1)$, we have*

$$(i) \quad \|F - \beta H\|_{\mathrm{MAX}} = O_p\left((\Delta_n \log d)^{1/2} + d^{-1/2}m_d\right). \tag{A.7}$$

$$(ii) \quad \left\|H^{-1}\right\| = O_p(1). \tag{A.8}$$

$$(iii) \quad \left\|G - H^{-1}\mathcal{X}\right\| = O_p\left((\Delta_n \log d)^{1/2} + d^{-1/2}m_d\right). \tag{A.9}$$

Proof of Lemma 2. (i) By simple calculations, we have

$$\begin{aligned}
F - \beta H =& t^{-1}\left(\mathcal{Y}\mathcal{Y}^\intercal - \beta\mathcal{X}\mathcal{X}^\intercal\beta^\intercal\right)F\Lambda^{-1} \\
=& t^{-1}\left(\beta\mathcal{X}\mathcal{Z}^\intercal F\Lambda^{-1} + \mathcal{Z}\mathcal{X}^\intercal\beta^\intercal F\Lambda^{-1} + (\mathcal{Z}\mathcal{Z}^\intercal - \Gamma)F\Lambda^{-1} + \Gamma F\Lambda^{-1}\right).
\end{aligned} \tag{A.10}$$

We bound these terms separately. First, we have

$$\left\|(\mathcal{Z}\mathcal{Z}^\intercal - \Gamma)F\Lambda^{-1}\right\|_{\mathrm{MAX}} \leq \|\mathcal{Z}\mathcal{Z}^\intercal - \Gamma\|_{\mathrm{MAX}} \|F\|_1 \left\|\Lambda^{-1}\right\|_{\mathrm{MAX}}.$$

Moroever, $\|F\|_1 \leq d^{1/2}\|F\|_{\mathrm{F}} = d$, and by (A.5), $\left\|\Lambda^{-1}\right\|_{\mathrm{MAX}} = O_p(d^{-1})$, which implies that

$$\left\|(\mathcal{Z}\mathcal{Z}^\intercal - \Gamma)F\Lambda^{-1}\right\|_{\mathrm{MAX}} = O_p((\Delta_n \log d)^{1/2}).$$

In addition, since $\|\Gamma\|_\infty \leq Km_d$ and $\|F\|_{\mathrm{MAX}} \leq \|F\|_{\mathrm{F}} = d^{1/2}$, it follows that

$$\left\|\Gamma F\Lambda^{-1}\right\|_{\mathrm{MAX}} \leq \|\Gamma\|_\infty \|F\|_{\mathrm{MAX}} \left\|\Lambda^{-1}\right\|_{\mathrm{MAX}} = O_p(d^{-1/2}m_d).$$

Also, we have

$$\left\|\beta\mathcal{X}\mathcal{Z}^\intercal F\Lambda^{-1}\right\|_{\mathrm{MAX}} \leq \|\beta\|_{\mathrm{MAX}} \|\mathcal{X}\mathcal{Z}^\intercal\|_1 \|F\|_1 \left\|\Lambda^{-1}\right\|_{\mathrm{MAX}} = O_p((\Delta_n \log d)^{1/2}).$$

where we use the fact that $\|\beta\|_{\mathrm{MAX}} \leq K$ and the bound below derived from (A.2):

$$\|\mathcal{X}\mathcal{Z}^\intercal\|_1 = \max_{1\leq l\leq d}\sum_{j=1}^{r}\left|\sum_{i=1}^{n}(\Delta_i^n X_j)(\Delta_i^n Z_l)\right| \leq r \max_{1\leq l\leq d, 1\leq j\leq r}\left|\sum_{i=1}^{n}(\Delta_i^n X_j)(\Delta_i^n Z_l)\right| = O_p((\Delta_n \log d)^{1/2}).$$

The remainder term can be bounded similarly.

(ii) Since $\|\beta\| = O(d^{1/2})$ and $\left\|t^{-1}\mathcal{X}\mathcal{X}^\intercal\right\| = O_p(1)$, we have

$$\|H\| = \left\|t^{-1}\mathcal{X}\mathcal{X}^\intercal\beta^\intercal F\Lambda^{-1}\right\| \leq \left\|t^{-1}\mathcal{X}\mathcal{X}^\intercal\right\| \|\beta\| \|F\| \left\|\Lambda^{-1}\right\| = O_p(1).$$

By triangle inequalities, and that $\|F - \beta H\| \leq (rd)^{1/2}\|F - \beta H\|_{\mathrm{MAX}}$, we have

$$\begin{aligned}
\|H^\intercal H - \mathrm{I}_r\| \leq& \left\|H^\intercal H - d^{-1}H^\intercal\beta^\intercal\beta H\right\| + d^{-1}\|H^\intercal\beta^\intercal\beta H - \mathrm{I}_r\| \\
\leq& \|H\|^2\left\|\mathrm{I}_r - d^{-1}\beta^\intercal\beta\right\| + d^{-1}\|H^\intercal\beta^\intercal\beta H - F^\intercal F\| \\
\leq& \|H\|^2\left\|\mathrm{I}_r - d^{-1}\beta^\intercal\beta\right\| + d^{-1}\|F - \beta H\|\|\beta H\| + d^{-1}\|F - \beta H\|\|F\| \\
=& o_p(1).
\end{aligned}$$

By Weyl's inequality again, we have $\lambda_{\min}(H^\intercal H) > 1/2$ with probability approaching 1. Therefore, H is invertible, and $\left\|H^{-1}\right\| = O_p(1)$.

(iii) We use the following decomposition:

$$G - H^{-1}\mathcal{X} = d^{-1}F^\intercal(\beta H - F)H^{-1}\mathcal{X} + d^{-1}(F^\intercal - H^\intercal\beta^\intercal)\mathcal{Z} + d^{-1}H^\intercal\beta^\intercal\mathcal{Z}.$$

Note that by (A.4), we have $\|\mathcal{X}\| = O_p(1)$. Moreover, since $\|F\| \leq \|F\|_{\mathrm{F}}$ and $\|F - \beta H\| \leq d^{1/2}\|F - \beta H\|_{\mathrm{MAX}}$, we have

$$\left\|d^{-1}F^\intercal(\beta H - F)H^{-1}\mathcal{X}\right\| \leq d^{-1}\|F\|\|F - \beta H\|\left\|H^{-1}\right\|\|\mathcal{X}\| = O_p\left((\Delta_n \log d)^{1/2} + d^{-1/2}m_d\right).$$

Similarly, by (A.6) we have

$$\|\mathcal{Z}\| = O_p(d^{1/2}(\Delta_n \log d)^{1/4} + m_d^{1/2}),$$

which leads to

$$\left\|d^{-1}(F^\intercal - H^\intercal\beta^\intercal)\mathcal{Z}\right\| = O_p\left(\left((\Delta_n \log d)^{1/4} + d^{-1/2}m_d^{1/2}\right)\left((\Delta_n \log d)^{1/2} + d^{-1/2}m_d\right)\right).$$

Moreover, we can apply Lemma 1 to $\beta^\intercal \mathcal{Z}$, which is an $r \times n$ matrix, so we have

$$\begin{aligned}\|\beta^\intercal \mathcal{Z}\| =& \sqrt{\lambda_1\left(\beta^\intercal \mathcal{Z}\mathcal{Z}^\intercal \beta\right)} \leq \sqrt{\|\beta^\intercal \mathcal{Z}\mathcal{Z}^\intercal \beta - \beta^\intercal \Gamma \beta\|_\infty + \|\beta\|_1 \|\Gamma\|_\infty \|\beta\|_\infty} \\ \leq& K\left(\Delta_n \log d\right)^{1/4} + K d^{1/2} m_d^{1/2},\end{aligned}$$

which leads to

$$\left\|d^{-1} H^\intercal \beta^\intercal \mathcal{Z}\right\| = O_p\left(d^{-1}(\Delta_n \log d)^{1/4} + d^{-1/2} m_d^{1/2}\right).$$

This concludes the proof. □

Lemma 3. *Under Assumptions 1 - 4, $d^{-1/2} m_d = o(1)$, and $\Delta_n \log d = o(1)$, we have*

$$\left\|\widehat{\Gamma}^S - \Gamma\right\|_{\text{MAX}} \leq \left\|\widehat{\Gamma} - \Gamma\right\|_{\text{MAX}} = O_p\left((\Delta_n \log d)^{1/2} + d^{-1/2} m_d\right). \tag{A.11}$$

Proof of Lemma 3. We write $G = (g_1, g_2, \ldots, g_n)$, $F = (f_1, f_2, \ldots, f_d)^\intercal$, $\beta = (\beta_1, \beta_2, \ldots, \beta_d)^\intercal$, and $\widehat{\Delta_i^n Z_k} = \Delta_i^n Y_k - f_k^\intercal g_i$. Hence, $\widehat{\Gamma}_{lk} = t^{-1} \sum_{i=1}^n (\widehat{\Delta_i^n Z_l})(\widehat{\Delta_i^n Z_k})$.

For $1 \leq k \leq d$ and $1 \leq i \leq n$, we have

$$\begin{aligned}\Delta_i^n Z_k - \widehat{\Delta_i^n Z_k} =& \Delta_i^n Y_k - \beta_k^\intercal \Delta_i^n X - (\Delta_i^n Y_k - f_k^\intercal g_i) = f_k^\intercal g_i - \beta_k^\intercal \Delta_i^n X \\ =& \beta_k^\intercal H(g_i - H^{-1}\Delta_i^n X) + (f_k^\intercal - \beta_k^\intercal H)(g_i - H^{-1}\Delta_i^n X) + (f_k^\intercal - \beta_k^\intercal H) H^{-1} \Delta_i^n X.\end{aligned}$$

Therefore, using $(a+b+c)^2 \leq 3(a^2+b^2+c^2)$, we have

$$\begin{aligned}&\sum_{i=1}^n \left(\Delta_i^n Z_k - \widehat{\Delta_i^n Z_k}\right)^2 \\ \leq& 3\sum_{i=1}^n \left(\beta_k^\intercal H(g_i - H^{-1}\Delta_i^n X)\right)^2 + 3\sum_{i=1}^n \left((f_k^\intercal - \beta_k^\intercal H)(g_i - H^{-1}\Delta_i^n X)\right)^2 + 3\sum_{i=1}^n \left((f_k^\intercal - \beta_k^\intercal H) H^{-1}\Delta_i^n X\right)^2.\end{aligned}$$

Using $v^\intercal A v \leq \lambda_{\max}(A) v^\intercal v$ repeatedly, if follows that

$$\begin{aligned}\sum_{i=1}^n \left(\beta_k^\intercal H(g_i - H^{-1}\Delta_i^n X)\right)^2 =& \sum_{i=1}^n \beta_k^\intercal H(G - H^{-1}\mathcal{X}) e_i e_i^\intercal (G - H^{-1}\mathcal{X})^\intercal H^\intercal \beta_k \\ \leq& \lambda_{\max}\left(\left(G - H^{-1}\mathcal{X}\right)\left(\left(G - H^{-1}\mathcal{X}\right)^\intercal\right)\lambda_{\max}(HH^\intercal)\beta_k^\intercal \beta_k \\ \leq& r\left\|G - H^{-1}\mathcal{X}\right\|^2 \|H\|^2 \max_{1\leq l\leq r} |\beta_{kl}|^2\end{aligned}$$

Similarly, we can bound the other terms.

$$\begin{aligned}\sum_{i=1}^n \left((f_k^\intercal - \beta_k^\intercal H)(g_i - H^{-1}\Delta_i^n X)\right)^2 \leq& r\left\|G - H^{-1}\mathcal{X}\right\|^2 \max_{1\leq l\leq r} (F_{kl} - (\beta_k^\intercal H)_l)^2, \\ \sum_{i=1}^n \left((f_k^\intercal - \beta_k^\intercal H) H^{-1}\Delta_i^n X\right)^2 \leq& rt\|E\|\left\|H^{-1}\right\|^2 \max_{1\leq l\leq r} (F_{kl} - (\beta_k^\intercal H)_l)^2.\end{aligned}$$

As a result, by Lemma 2, we have

$$\begin{aligned}
&\max_{1\le k\le d}\sum_{i=1}^{n}\left(\Delta_i^n Z_k-\widehat{\Delta_i^n Z_k}\right)^2\\
\le& K\left\|G-H^{-1}\mathcal{X}\right\|^2\|H\|^2\|\beta\|_{\max}^2+K\left\|G-H^{-1}\mathcal{X}\right\|^2\|F-\beta H\|_{\max}^2+K\|E\|\left\|H^{-1}\right\|^2\|F-\beta H\|_{\max}^2\\
\le& O_p\left((\Delta_n\log d)+d^{-1}m_d^2\right)
\end{aligned}$$

By the Cauchy-Schwartz inequality, we have

$$\begin{aligned}
&\max_{1\le l,k\le d}\left|\sum_{i=1}^{n}(\widehat{\Delta_i^n Z_l})(\widehat{\Delta_i^n Z_k})-\sum_{i=1}^{n}(\Delta_i^n Z_l)(\Delta_i^n Z_k)\right|\\
\le&\max_{1\le l,k\le d}\left|\sum_{i=1}^{n}\left(\widehat{\Delta_i^n Z_l}-\Delta_i^n Z_l\right)\left(\widehat{\Delta_i^n Z_k}-\Delta_i^n Z_k\right)\right|+2\max_{1\le l,k\le d}\left|\sum_{i=1}^{n}(\Delta_i^n Z_l)\left(\widehat{\Delta_i^n Z_k}-\Delta_i^n Z_k\right)\right|\\
\le&\max_{1\le l\le d}\sum_{i=1}^{n}\left(\widehat{\Delta_i^n Z_l}-\Delta_i^n Z_l\right)^2+2\sqrt{\max_{1\le l\le d}\sum_{i=1}^{n}(\Delta_i^n Z_l)^2\max_{1\le l\le d}\sum_{i=1}^{n}\left(\widehat{\Delta_i^n Z_l}-\Delta_i^n Z_l\right)^2}\\
=&O_p\left((\Delta_n\log d)^{1/2}+d^{-1/2}m_d\right),
\end{aligned}$$

Finally, by the triangular inequality,

$$\begin{aligned}
\max_{1\le l,k\le d,(l,k)\in S}\left|\widehat{\Gamma}_{lk}-\Gamma_{lk}\right|\le&\max_{1\le l,k\le d}\left|\widehat{\Gamma}_{lk}-\Gamma_{lk}\right|\le\max_{1\le l,k\le d}\left|\sum_{i=1}^{n}(\Delta_i^n Z_l)(\Delta_i^n Z_k)-\int_0^t g_{s,lk}ds\right|\\
&+\max_{1\le l,k\le d}\left|\sum_{i=1}^{n}(\widehat{\Delta_i^n Z_l})(\widehat{\Delta_i^n Z_k})-\sum_{i=1}^{n}(\Delta_i^n Z_l)(\Delta_i^n Z_k)\right|,
\end{aligned}$$

which yields the desired result by using (A.1). □

Lemma 4. *Under Assumptions 1 - 4, $d^{-1/2}m_d=o(1)$, and $\Delta_n\log d=o(1)$, we have*

$$\left\|t^{-1}FGG^{\intercal}F^{\intercal}-\beta\mathrm{E}\beta^{\intercal}\right\|_{\max}=O_p\left((\Delta_n\log d)^{1/2}+d^{-1/2}m_d\right).$$

Proof. By simple calculations, we have

$$\begin{aligned}
\left\|t^{-1}FGG^{\intercal}F-\beta\mathrm{E}\beta^{\intercal}\right\|_{\max}\le&d^{-1}\left\|(F-\beta H)\Lambda(F-\beta H)^{\intercal}\right\|_{\max}+2d^{-1}\left\|\beta H\Lambda(F-\beta H)^{\intercal}\right\|_{\max}\\
&+t^{-1}\left\|\beta H(G-H^{-1}\mathcal{X})(G-H^{-1}\mathcal{X})^{\intercal}H^{\intercal}\beta^{\intercal}\right\|_{\max}\\
&+2t^{-1}\left\|\beta H(G-H^{-1}\mathcal{X})\mathcal{X}^{\intercal}\beta^{\intercal}\right\|_{\max}+\left\|\beta(t^{-1}\mathcal{X}\mathcal{X}^{\intercal}-\mathrm{E})\beta^{\intercal}\right\|_{\max}
\end{aligned}$$

Note that by Lemma 2, (A.4), $\|\beta\|_{\max}=O_p(1)$, $\|H\|=O_p(1)$, and $\|\Lambda\|_{\max}=O_p(1)$,

$$\begin{aligned}
d^{-1}\left\|(F-\beta H)\Lambda(F-\beta H)^{\intercal}\right\|_{\max}\le&r^2d^{-1}\|F-\beta H\|_{\max}^2\|\Lambda\|_{\max}\\
\le&O_p(\Delta_n\log d+d^{-1}m_d^2),\\
2d^{-1}\left\|\beta H\Lambda(F-\beta H)^{\intercal}\right\|_{\max}\le&2r^2d^{-1}\|\beta\|_{\max}\|H\|\|\Lambda\|_{\max}\|F-\beta H\|_{\max}
\end{aligned}$$

$$
\begin{aligned}
&\leq O_p\left((\Delta_n \log d)^{1/2} + d^{-1/2} m_d\right), \\
t^{-1}\left\|\beta H (G - H^{-1}\mathcal{X})(G - H^{-1}\mathcal{X})^\intercal H^\intercal \beta^\intercal\right\|_{\text{MAX}} &\leq r^4 t^{-1} \|\beta\|_{\text{MAX}}^2 \|H\|^2 \left\|G - H^{-1}\mathcal{X}\right\|^2 \\
&\leq O_p(\Delta_n \log d + d^{-1} m_d^2), \\
2t^{-1}\left\|\beta H (G - H^{-1}\mathcal{X})\mathcal{X}^\intercal \beta^\intercal\right\|_{\text{MAX}} &\leq r^3 \|\beta\|_{\text{MAX}}^2 \|H\| \left\|G - H^{-1}\mathcal{X}\right\| \|\mathcal{X}\| \\
&\leq O_p\left((\Delta_n \log d)^{1/2} + d^{-1/2} m_d\right), \\
\left\|\beta(t^{-1}\mathcal{X}\mathcal{X}^\intercal - \mathrm{E})\beta^\intercal\right\|_{\text{MAX}} &\leq r^2 \|\beta\|_{\text{MAX}} \left\|t^{-1}\mathcal{X}\mathcal{X}^\intercal - \mathrm{E}\right\|_{\text{MAX}} \\
&\leq O_p\left((\Delta_n \log d)^{1/2}\right).
\end{aligned}
$$

Combining the above inequalities concludes the proof. □

Proof of Theorem 3. Note that

$$
\widehat{\Sigma}^S = d^{-1} F \Lambda F^\intercal + \widehat{\Gamma}^S = t^{-1} F G G^\intercal F^\intercal + \widehat{\Gamma}^S.
$$

By Lemma 3, we have

$$
\left\|\widehat{\Gamma}^S - \Gamma\right\|_{\text{MAX}} = O_p\left((\Delta_n \log d)^{1/2} + d^{-1/2} m_d\right).
$$

By the triangle inequality, we have

$$
\left\|\widehat{\Sigma}^S - \Sigma\right\|_{\text{MAX}} \leq \left\|d^{-1} F \Lambda F^\intercal - \beta \mathrm{E} \beta^\intercal\right\|_{\text{MAX}} + \left\|\widehat{\Gamma}^S - \Gamma\right\|_{\text{MAX}}
$$

Therefore, the desired result follows from Lemmas 3 and 4. □

Appendix A.4 Proof of Theorem 4

Lemma 5. *Under Assumptions 1 - 4, $d^{-1/2} m_d = o(1)$, and $\Delta_n \log d = o(1)$, we have*

$$
\left\|\widehat{\Gamma}^S - \Gamma\right\| = O_p\left(m_d (\Delta_n \log d)^{1/2} + d^{-1/2} m_d^2\right). \tag{A.12}
$$

Moreover, if in addition, $d^{-1/2} m_d^2 = o(1)$ and $m_d (\Delta_n \log d)^{1/2} = o(1)$ hold, then $\lambda_{\min}\left(\widehat{\Gamma}^S\right)$ is bounded away from 0 with probability approaching 1, and

$$
\left\|\left(\widehat{\Gamma}^S\right)^{-1} - \Gamma^{-1}\right\| = O_p\left(m_d (\Delta_n \log d)^{1/2} + d^{-1/2} m_d^2\right).
$$

Proof of Lemma 5. Note that since $\widehat{\Gamma}^S - \Gamma$ is symmetric,

$$
\left\|\widehat{\Gamma}^S - \Gamma\right\| \leq \left\|\widehat{\Gamma}^S - \Gamma\right\|_\infty = \max_{1 \leq l \leq d} \sum_{k=1}^{d} \left|\widehat{\Gamma}_{lk}^S - \Gamma_{lk}\right| \leq m_d \max_{1 \leq l \leq d, 1 \leq k \leq d} \left|\widehat{\Gamma}_{lk}^S - \Gamma_{lk}\right|
$$

By Lemma 3, we have

$$
\left\|\widehat{\Gamma}^S - \Gamma\right\| \leq m_d \left\|\widehat{\Gamma}^S - \Gamma\right\|_{\text{MAX}} = O_p\left(m_d (\Delta_n \log d)^{1/2} + d^{-1/2} m_d^2\right).
$$

Moreover, since $\lambda_{\min}(\Gamma) > K$ for some constant K and by Weyl's inequality, we have $\lambda_{\min}(\widehat{\Gamma}^S) > K - o_p(1)$. As a result, we have

$$\begin{aligned}\left\|\left(\widehat{\Gamma}^S\right)^{-1} - \Gamma^{-1}\right\| =& \left\|\left(\widehat{\Gamma}^S\right)^{-1}\left(\Gamma - \left(\widehat{\Gamma}^S\right)\right)\Gamma^{-1}\right\| \leq \lambda_{\min}(\widehat{\Gamma}^S)^{-1}\lambda_{\min}(\Gamma)^{-1}\left\|\Gamma - \widehat{\Gamma}^S\right\| \\ \leq& O_p\left(m_d(\Delta_n \log d)^{1/2} + d^{-1/2}m_d^2\right).\end{aligned}$$

□

Proof of Theorem 4. First, by Lemma 5 and the fact that $\lambda_{\min}(\widehat{\Sigma}^S) \geq \lambda_{\min}(\widehat{\Gamma}^S)$, we can establish the first two statements.

To bound $\left\|(\widehat{\Sigma}^S)^{-1} - \Sigma^{-1}\right\|$, by the Sherman - Morrison - Woodbury formula, we have

$$\begin{aligned}&\left(\widehat{\Sigma}^S\right)^{-1} - \left(\widetilde{\Sigma}\right)^{-1} \\ =&\left(t^{-1}FGG^\intercal F^\intercal + \widehat{\Gamma}^S\right)^{-1} - \left(t^{-1}\beta H H^{-1}\mathcal{X}\mathcal{X}^\intercal (H^{-1})^\intercal H^\intercal \beta^\intercal + \Gamma\right)^{-1} \\ =&\left((\widehat{\Gamma}^S)^{-1} - \Gamma^{-1}\right) - \left((\widehat{\Gamma}^S)^{-1} - \Gamma^{-1}\right) F\left(d\Lambda^{-1} + F^\intercal(\widehat{\Gamma}^S)^{-1}F\right)^{-1} F^\intercal(\widehat{\Gamma}^S)^{-1} \\ &- \Gamma^{-1}F\left(d\Lambda^{-1} + F^\intercal(\widehat{\Gamma}^S)^{-1}F\right)^{-1} F^\intercal\left((\widehat{\Gamma}^S)^{-1} - \Gamma^{-1}\right) \\ &+ \Gamma^{-1}(\beta H - F)\left(tH^\intercal(\mathcal{X}\mathcal{X}^\intercal)^{-1}H + H^\intercal\beta^\intercal\Gamma^{-1}\beta H\right)^{-1} H^\intercal\beta^\intercal\Gamma^{-1} \\ &- \Gamma^{-1}F\left(tH^\intercal(\mathcal{X}\mathcal{X}^\intercal)^{-1}H + H^\intercal\beta^\intercal\Gamma^{-1}\beta H\right)^{-1}(F^\intercal - H^\intercal\beta^\intercal)\Gamma^{-1} \\ &+ \Gamma^{-1}F\left(\left(tH^\intercal(\mathcal{X}\mathcal{X}^\intercal)^{-1}H + H^\intercal\beta^\intercal\Gamma^{-1}\beta H\right)^{-1} - \left(d\Lambda^{-1} + F^\intercal(\widehat{\Gamma}^S)^{-1}F\right)^{-1}\right)F^\intercal\Gamma^{-1} \\ =&L_1 + L_2 + L_3 + L_4 + L_5 + L_6.\end{aligned}$$

By Lemma 5, we have

$$\|L_1\| = O_p\left(m_d(\Delta_n \log d)^{1/2} + d^{-1/2}m_d^2\right).$$

For L_2, because $\|F\| = O_p(d^{1/2})$, $\lambda_{\max}\left((\widehat{\Gamma}^S)^{-1}\right) \leq \left(\lambda_{\min}(\widehat{\Gamma}^S)\right)^{-1} \leq K + o_p(1)$,

$$\lambda_{\min}\left(d\Lambda^{-1} + F^\intercal(\widehat{\Gamma}^S)^{-1}F\right) \geq \lambda_{\min}\left(F^\intercal(\widehat{\Gamma}^S)^{-1}F\right) \geq \lambda_{\min}\left(F^\intercal F\right)\lambda_{\min}\left((\widehat{\Gamma}^S)^{-1}\right) \geq m_d^{-1}d,$$

and by Lemma 5, we have

$$\begin{aligned}\|L_2\| \leq& \left\|\left((\widehat{\Gamma}^S)^{-1} - \Gamma^{-1}\right)\right\| \|F\| \left\|\left(d\Lambda^{-1} + F^\intercal(\widehat{\Gamma}^S)^{-1}F\right)^{-1}\right\| \left\|F^\intercal(\widehat{\Gamma}^S)^{-1}\right\| \\ =& O_p\left(m_d^2(\Delta_n \log d)^{1/2} + d^{-1/2}m_d^3\right).\end{aligned}$$

The same bound holds for $\|L_3\|$. As for L_4, note that $\|\beta\| = O_p(d^{1/2})$, $\|H\| = O_p(1)$, $\left\|\Gamma^{-1}\right\| \leq (\lambda_{\min}(\Gamma))^{-1} \leq K$, and $\|\beta H - F\| \leq \sqrt{rd}\,\|\beta H - F\|_{\text{MAX}} = O_p(d^{1/2}(\Delta_n \log d)^{1/2} + m_d)$, and that

$$\lambda_{\min}\left(tH^\intercal(\mathcal{X}\mathcal{X}^\intercal)^{-1}H + H^\intercal\beta^\intercal\Gamma^{-1}\beta H\right) \geq \lambda_{\min}\left(H^\intercal\beta^\intercal\Gamma^{-1}\beta H\right) \geq \lambda_{\min}(\Gamma^{-1})\lambda_{\min}(\beta^\intercal\beta)\lambda_{\min}(H^\intercal H)$$

$$> K m_d^{-1} d,$$

hence we have

$$\begin{aligned}\|L_4\| &\leq \left\|\Gamma^{-1}\right\| \|(\beta H - F)\| \left\|\left(tH^\intercal \left(\mathcal{X}\mathcal{X}^\intercal\right)^{-1} H + H^\intercal \beta^\intercal \Gamma^{-1} \beta H\right)^{-1}\right\| \|H^\intercal \beta^\intercal\| \left\|\Gamma^{-1}\right\| \\ &= O_p(m_d(\Delta_n \log d)^{1/2} + d^{-1/2} m_d^2).\end{aligned}$$

The same bound holds for L_5. Finally, with respect to L_6, we have

$$\begin{aligned}&\left\|\left(tH^\intercal \left(\mathcal{X}\mathcal{X}^\intercal\right)^{-1} H + H^\intercal \beta^\intercal \Gamma^{-1} \beta H\right)^{-1} - \left(d\Lambda^{-1} + F^\intercal (\widehat{\Gamma}^S)^{-1} F\right)^{-1}\right\| \\ \leq & K d^{-2} m_d^2 \left\|\left(tH^\intercal \left(\mathcal{X}\mathcal{X}^\intercal\right)^{-1} H + H^\intercal \beta^\intercal \Gamma^{-1} \beta H\right) - \left(d\Lambda^{-1} + F^\intercal (\widehat{\Gamma}^S)^{-1} F\right)\right\|.\end{aligned}$$

Moreover, since we have

$$\left\|tH^\intercal (\mathcal{X}\mathcal{X}^\intercal)^{-1} H - d\Lambda^{-1}\right\| = \left\|\Lambda^{-1} F^\intercal (\beta H - F)\right\| = O_p\left((\Delta_n \log d)^{1/2} + d^{-1/2} m_d\right)$$

and

$$\begin{aligned}&\left\|H^\intercal \beta^\intercal \Gamma^{-1} \beta H - F^\intercal (\widehat{\Gamma}^S)^{-1} F\right\| \\ \leq & \left\|(H^\intercal \beta^\intercal - F^\intercal)\Gamma^{-1} \beta H\right\| + \left\|F^\intercal \Gamma^{-1} (\beta H - F)\right\| + \left\|F^\intercal \left(\Gamma^{-1} - (\widehat{\Gamma}^S)^{-1}\right) F\right\| \\ = & O_p\left(d m_d (\Delta_n \log d)^{1/2} + d^{1/2} m_d^2\right),\end{aligned}$$

combining these inequalities yields

$$\|L_6\| = O_p\left(m_d^3 (\Delta_n \log d)^{1/2} + d^{-1/2} m_d^4\right).$$

On the other hand, using the Sherman - Morrison - Woodbury formula again,

$$\begin{aligned}&\left\|\widetilde{\Sigma}^{-1} - \Sigma^{-1}\right\| \\ = & \left\|\left(t^{-1} \beta \mathcal{X}\mathcal{X}^\intercal \beta^\intercal + \Gamma\right)^{-1} - \left(\beta \mathrm{E} \beta^\intercal + \Gamma\right)^{-1}\right\| \\ \leq & \left\|\Gamma^{-1}\right\|^2 \|\beta H\|^2 \left\|\left(\left(tH^\intercal \left(\mathcal{X}\mathcal{X}^\intercal\right)^{-1} H + H^\intercal \beta^\intercal \Gamma^{-1} \beta H\right)^{-1} - \left(H^\intercal \mathrm{E}^{-1} H + H^\intercal \beta^\intercal \Gamma^{-1} \beta H\right)^{-1}\right)\right\| \\ \leq & K d \left\|tH^\intercal \left(\mathcal{X}\mathcal{X}^\intercal\right)^{-1} H + H^\intercal \beta^\intercal \Gamma^{-1} \beta H\right\|^{-1} \left\|H^\intercal \mathrm{E}^{-1} H + H^\intercal \beta^\intercal \Gamma^{-1} \beta H\right\|^{-1} \left\|t\left(\mathcal{X}\mathcal{X}^\intercal\right)^{-1} - \mathrm{E}^{-1}\right\| \\ = & O_p\left(m_d (\Delta_n \log d)^{1/2}\right).\end{aligned}$$

By the triangle inequality, we obtain

$$\left\|(\widehat{\Sigma}^S)^{-1} - \Sigma^{-1}\right\| \leq \left\|(\widehat{\Sigma}^S)^{-1} - \widetilde{\Sigma}^{-1}\right\| + \left\|\widetilde{\Sigma}^{-1} - \Sigma^{-1}\right\| = O_p\left(m_d^3 (\Delta_n \log d)^{1/2} + d^{-1/2} m_d^4\right).$$

□

Appendix A.5 Proof of Theorem 5

Proof of Theorem 5. This have been established by Lemma 2. □